

DIREITO AUTORAL E APRENDIZAGEM DE IA GENERATIVA: ANÁLISE COMPARADA BRASIL E ESTADOS UNIDOS NO USO DE OBRAS PROTEGIDAS PARA TREINAMENTO DE IA

COPYRIGHT LAW AND GENERATIVE ARTIFICIAL INTELLIGENCE: A
COMPARATIVE ANALYSIS OF BRAZIL AND THE UNITED STATES IN THE USE
OF PROTECTED WORKS FOR AI TRAINING

DERECHO DE AUTOR Y APRENDIZAJE DE IA GENERATIVA: ANÁLISIS
COMPARADO ENTRE BRASIL Y ESTADOS UNIDOS SOBRE EL USO DE OBRAS
PROTEGIDAS PARA EL ENTRENAMIENTO DE IA

Michele Cristine Soares Campos Volpe¹

DOI: 10.54899/dcs.v22i84.3880

Recibido: 24/10/2025 | Aceptado: 21/11/2025 | Publicación en línea: 26/11/2025.

RESUMO

O artigo analisa os desafios jurídicos decorrentes do uso de obras protegidas por direito autoral no treinamento de sistemas de Inteligência Artificial Generativa (IAG). Examinando como o avanço de modelos como ChatGPT, Gemini e DALL-E, dependentes de grandes volumes de dados obtidos por *web scraping*, tensiona os limites tradicionais da legislação autoral. O estudo adota abordagem jurídico-comparada entre os sistemas dos Estados Unidos e do Brasil, destacando que, enquanto o ordenamento norte-americano admite a aplicação da doutrina do *fair use* - cuja análise se baseia nos fatores de propósito, natureza, quantidade e impacto de mercado -, o direito brasileiro, de matriz personalista, não comporta tal exceção, exigindo autorização prévia e expressa para qualquer uso de obras protegidas. O caso Folha de S. Paulo v. OpenAI é apontado como primeiro grande teste da LDA frente à IAG, evidenciando as lacunas regulatórias nacionais. O trabalho também discute a impossibilidade técnica de mensuração de *royalties*, em razão da natureza probabilística dos modelos, e a necessidade de harmonização normativa por meio de transparência e compensação coletiva, conforme proposto no Projeto de Lei nº 2.338/2023. Conclui-se que a conciliação entre inovação tecnológica e tutela autoral demanda um arcabouço jurídico multidisciplinar que assegure equilíbrio entre desenvolvimento e proteção da criação humana.

Palavras-chave: Inteligência Artificial Generativa. Direito Autoral. Fair Use. Web Scraping. Mineração de Dados.

ABSTRACT

The article analyzes the legal challenges arising from the use of copyright-protected works in the training of Generative Artificial Intelligence (GAI) systems. It examines how the advancement of models such as ChatGPT, Gemini, and DALL-E - dependent on large volumes of data obtained

¹ Mestranda em Direito Comercial pela Pontifícia Universidade Católica de São Paulo (PUC - SP), São Paulo, São Paulo, Brasil. E-mail: michele.volpe@camposvolpe.com.br Orcid: <https://orcid.org/0009-0002-4659-4310>

through web scraping - stretches the traditional boundaries of copyright law. The study adopts a comparative legal approach between the United States and Brazil, emphasizing that while the U.S. legal system admits the application of the *fair use* doctrine - whose analysis is based on the factors of purpose, nature, amount, and market impact - the Brazilian system, rooted in a personalist tradition, does not recognize such an exception, requiring prior and express authorization for any use of protected works. The *Folha de S. Paulo v. OpenAI* case is identified as the first major test of Brazil's Copyright Law (LDA) in relation to GAI, highlighting national regulatory gaps. The paper also discusses the technical impossibility of calculating royalties due to the probabilistic nature of the models, as well as the need for normative harmonization through transparency and collective compensation mechanisms, as proposed in Bill No. 2.338/2023. It concludes that reconciling technological innovation with copyright protection requires a multidisciplinary legal framework that ensures balance between development and the safeguarding of human creativity.

Keywords: Generative Artificial Intelligence. Copyright Law. Fair Use. Web Scraping. Data Mining.

RESUMEN

El artículo analiza los desafíos jurídicos derivados del uso de obras protegidas por derechos de autor en el entrenamiento de sistemas de Inteligencia Artificial Generativa (IAG). Examina cómo el avance de modelos como ChatGPT, Gemini y DALL-E, que dependen de grandes volúmenes de datos obtenidos mediante web scraping, tensiona los límites tradicionales de la legislación autoral. El estudio adopta un enfoque jurídico-comparado entre los sistemas de Estados Unidos y Brasil, destacando que, mientras el ordenamiento estadounidense admite la aplicación de la doctrina del *fair use* —cuyo análisis se basa en los factores de propósito, naturaleza, cantidad e impacto en el mercado—, el derecho brasileño, de matriz personalista, no contempla tal excepción, exigiendo autorización previa y expresa para cualquier uso de obras protegidas. El caso *Folha de S. Paulo v. OpenAI* se señala como la primera gran prueba de la Ley de Derechos de Autor frente a la IAG, evidenciando las lagunas regulatorias nacionales. El trabajo también discute la imposibilidad técnica de medir regalías debido a la naturaleza probabilística de los modelos, así como la necesidad de una armonización normativa basada en la transparencia y la compensación colectiva, conforme a lo propuesto en el Proyecto de Ley n.º 2.338/2023. Se concluye que la conciliación entre innovación tecnológica y tutela autoral exige un marco jurídico multidisciplinario que garantice el equilibrio entre el desarrollo y la protección de la creación humana.

Palabras clave: Inteligencia Artificial Generativa. Derecho de Autor. Fair Use. Web Scraping. Minería de Datos.



Esta obra está bajo una [Licencia Creative Commons Atribución- NoComercial 4.0 Internacional](https://creativecommons.org/licenses/by-nc/4.0/)

INTRODUÇÃO

O avanço acelerado da inteligência artificial generativa (IAG) inaugura um campo de tensões normativas que desafiam os limites tradicionais do direito autoral. Se, por um lado, esses sistemas se consolidam como motores de inovação, dependentes de vastos volumes de dados, por outro, a forma como tais dados são coletados e reutilizados - muitas vezes de obras protegidas por direitos autorais - suscita questionamentos centrais sobre licitude, remuneração e segurança jurídica. A questão, portanto, não é apenas tecnológica, mas jurídica e política, exigindo análise comparada das diferentes respostas normativas em construção.

O surgimento de Large Language Models (LLMs) como ChatGPT, Gemini, Deepseek e Perplexity, bem como de sistemas de geração de imagens como DALL-E e Midjourney, evidenciam uma dependência crítica de vastos volumes de dados para o treinamento desses sistemas, frequentemente obtidos através de técnicas de web scraping de conteúdo disponível na internet, o qual consiste em raspagem de conteúdos disponíveis em páginas web como insumo ao processamento de aprendizado de máquina.

Esta realidade tecnológica tem gerado tensões jurídicas sem precedentes entre os interesses de empresas desenvolvedoras de IAG - como OpenAI, Anthropic e Meta - e os titulares de direitos autorais, cujas obras são incorporadas aos datasets de treinamento sem autorização expressa.

A economia política dos dados revela um cenário complexo, no qual grandes corporações tecnológicas negociam licenças com editoras tradicionais enquanto simultaneamente utilizam conteúdo não licenciado obtido através de métodos automatizados de coleta.

Nesse contexto, o uso de datasets compostos por obras literárias, artísticas e científicas coloca em evidência uma problemática central: até que ponto o treinamento de IAG configura uso legítimo de obras preexistentes ou se aproxima de uma reprodução ou criação derivada não autorizada? A resposta a essa pergunta define não apenas os direitos dos autores, mas também a viabilidade de modelos de negócio que sustentam a economia da IAG.

As respostas jurídicas variam entre jurisdições, mas convergem na percepção de que o problema exige novas soluções. Nos Estados Unidos, litígios recentes exploram a aplicação da doutrina do *fair use* como justificativa para uso de obras no treinamento de IAG, enquanto na União Europeia o AI Act de 2024 impõem maior transparência e mecanismos de opt-out para titulares. Já no Brasil, o Projeto de Lei nº 2.338/2023, em tramitação na Câmara, avança no debate

ao prever hipóteses de remuneração dos autores pelo uso comercial de suas obras. Esse mosaico regulatório evidencia tanto divergências quanto pontos de interseção que podem orientar uma solução híbrida.

Está problemática conecta-se diretamente com questões mais amplas sobre a proteção da arquitetura de IAG, particularmente no que se refere à curadoria de dados como componente essencial dos sistemas neurais. A seleção, limpeza e organização de datasets constituem elementos críticos da arquitetura de IAG que enfrentam lacunas na proteção tradicional oferecida pelo sistema de patentes versus direito autoral.

Apesar de estarmos inseridos no contexto de uma sociedade altamente conectada produzindo dados a todo momento durante nosso uso dos mais diversos aparelhos tecnológicos, tais dados são ativos valiosos para tech companies, mas dispositivos de inteligência artificial necessitam de dados confiáveis, relevantes e com conteúdos diversificados para que possam ser adequadamente treinadas.

Atualmente manifesta-se crescimento exponencial de litígios envolvendo uso não autorizado de obras protegidas por direitos autorais no treinamento de IAG, como os casos paradigmáticos de The New York Times contra OpenAI e Microsoft, e o caso Authors Guild contra Google que fornecem precedentes importantes para análise de mineração de dados (text and data mining) em contextos tecnológicos.

Este estudo tem como objetivo analisar comparativamente as respostas do Estados Unidos e Brasil, a luz de suas legislações, ao uso de obras protegidas em datasets de treinamento de IAG, busca-se identificar os desafios comuns, avaliar os riscos de insegurança jurídica diante da evolução da tecnologia e proteção efetiva da propriedade intelectual.

Assim, o presente estudo adota metodologia jurídico-comparada, combinando análise normativa e documental de tratados internacionais, legislações e decisões judiciais relevantes, complementada pelo exame do caso Folha de S. Paulo v. OpenAI como estudo de caso emblemático da aplicação prática do direito autoral no contexto da inteligência artificial generativa.

CONCEITOS FUNDAMENTAIS DE INTELIGENCIA ARTIFICIAL E DADOS DE TREINAMENTO

Análise Gerais sobre Inteligência Artificial Generativa

A inteligência artificial insere-se na sociedade contemporâneo como um vetor de transformação profunda, influenciando não apenas os modos de produção econômica, mas também as estruturas sociais, éticas e jurídicas que regem as interações humanas, afinal, como afirma a professora Dora Kaufman (2022), a inteligência artificial pode ser considerada uma tecnologia de propósito geral do século XXI, sendo uma ferramenta poderosa que certamente mudará os contornos da sociedade dada sua vasta aplicabilidade.

A inteligência artificial generativa posiciona-se como instrumento catalisador de inovações que visam à resolução de problemas complexos, tais como otimização de processos industriais, avanços na saúde e aprimoramento da governança, ao mesmo tempo em que suscita debates sobre desigualdades digitais, empregabilidade e uso ético da ferramenta a luz de paradigmas sociais que não vislumbravam a velocidade de transformação vivenciada na nova era digital.

De acordo com perspectivas históricas, a inteligência artificial remonta aos anos 1950, com pioneiros como Alan Turing e John McCarthy, que delinearam os fundamentos teóricos para sistemas capazes de simular inteligência humana, evoluindo rapidamente de abordagens simbólicas para modelos probabilísticos e, mais recentemente, para arquiteturas baseadas em aprendizado profundo (deep learning), impulsionadas pelo aumento exponencial da capacidade computacional e da disponibilidade de dados massivos.

Nesta trajetória, a sociedade contemporânea testemunha a proliferação de aplicações de inteligência artificial em esferas diversas, desde assistentes virtuais em dispositivos cotidianos até algoritmos preditivos em sistemas financeiros e judiciais, fomentando uma economia digital estimada em trilhões de dólares.

No entanto, possui uma dualidade inerente: por um lado, promove inclusão ao democratizar acesso ao conhecimento e serviços de forma rápida e acessível por meio da internet, e por outro, agrava a concentração de poder nas gigantes corporações transnacionais detentoras de infraestruturas computacionais avançadas, como Google, OpenAI e Microsoft, que dominam o ecossistema global de desenvolvimento de inteligência artificial.

Dentre as manifestações mais proeminentes da inteligência artificial, destaca-se a inteligência artificial generativa (IAG), que transcende as capacidades analíticas tradicionais para englobar a criação autônoma de conteúdo, simulando processos criativos humanos, por deep learning através de métodos matemáticos probabilísticos, ou seja, a IAG é um sistema computacional que utiliza-se de arquitetura que visa simular o processo neuronal cerebral, cujos “neurônios artificiais” são treinado utilizando uma vasta gama de dados (como textos, imagens, áudios ou código de programação) de forma que ajusta pesos para aprender padrões estatísticos por método matemático decorrentes de padrões preexistentes entre tais dados, permitindo que a geração de novos conteúdos através de respostas (outputs) multimodais (textos, imagens ou áudios novos) de acordo com a solicitação (prompts) do usuário do sistema.

A IAG depende fortemente de grandes volumes de dados para funcionar de maneira eficaz, pois seus modelos precisam ser expostos a uma ampla diversidade de informações durante o treinamento, a fim de reconhecer padrões, contextos e relações complexas entre elementos. Quanto maior e mais variado for o conjunto de dados, maior tende a ser a capacidade da IAG de produzir resultados coerentes, criativos e adaptáveis a diferentes situações, e essa dependência significa que a qualidade e a representatividade dos dados utilizados impactam diretamente o desempenho do sistema, podendo gerar vieses ou limitações, caso os dados sejam insuficientes ou pouco diversificados.

Exemplos paradigmáticos incluem modelos como o GPT-4 da OpenAI, capaz de produzir narrativas coesas ou análises jurídicas preliminares, e o DALL-E, que sintetiza imagens visuais a partir de descrições textuais. Ambos, entretanto, ilustram o ponto de fricção entre tecnologia e direito: sua eficácia criativa é inseparável do acesso a bancos massivos de obras preexistentes, muitas delas sob proteção legal. Essa característica distingue a IAG como um objeto jurídico sui generis, pois coloca em debate se o processo de treinamento constitui mera utilização técnica de insumos ou uma forma de reprodução e/ou transformação de obras intelectuais.

E neste contexto que a problemática envolvendo os dados inicia-se, afinal a íntima dependência da tecnologia com os dados implica analisarmos que, tais informações, somente podem ser criadas de 2 (duas) formas: primariamente pelo autor principal - até pouco tempo único autor possível - o ser humano, ou secundariamente, a própria IAG produzindo conteúdo, os denominados dados sintéticos.

O treinamento da IAG requer a alimentação do sistema com muitos dados, bilhões de

tokens² derivados de textos, imagens ou músicas, de modo que os limites do direito autoral passam a ser testados não apenas no resultado gerado, mas também no processo de construção do modelo quanto a legitimidade do uso de obras humanas no momento do aprendizado algorítmico.

Assim hodiernamente IAs são criadas utilizando-se, preponderantemente, do produto da mente humana, no entanto tais autores não foram consultados previamente, informados sobre o uso e tampouco remunerados pelo consumo de suas obras, o que torna a IAG ainda mais complexa, não podendo ser compreendida apenas como evolução técnica, mas como um fenômeno jurídico-social dependente de dados protegidos por legislação específica, de forma que a análise de seus fundamentos técnicos, em especial a forma como a coletânea de dados é realizada, curada e integrada ao processo de aprendizado de máquina, é indispensável para compreender as controvérsias atuais em matéria de direito autoral.

Fundamentos Técnicos do Aprendizado de Máquina e uso de Data Scraping

A IAG contemporânea fundamenta-se, em sua forma mais difundida, no paradigma do aprendizado de máquina (machine learning), que representa uma abordagem computacional na qual sistemas desenvolvem capacidades de reconhecimento de padrões e tomada de decisões através da análise de grandes volumes de dados, dispensando programação explícita para cada tarefa específica.

Segundo Russell e Novig (2020), o processo de aprendizado de máquina (gênero) opera através de algoritmos matemáticos que identificam correlações estatísticas em conjuntos de dados estruturados ou não, ajustando progressivamente parâmetros internos para otimizar a precisão de suas previsões ou classificações. No contexto das redes neurais artificiais - “espécie” de aprendizado de máquina conhecido como deep learning - constituem a base dos modelos de linguagem (como ChatGPT) e geração de imagens mais avançados (a exemplo Midjourney), este processo envolve a calibração de milhões ou bilhões de pesos sinápticos artificiais através de técnicas de programação.

Modelos como GPT-4, Claude e Gemini requerem coleções de dados (denominadas datasets) contendo centenas de bilhões de parâmetros, a exemplo palavras, extraídas de fontes

² ² Segundo Nvidia, os tokens “são pequenas unidades de dados que vêm da quebra de pedaços maiores de informações. Os modelos de IA processam tokens para aprender as relações entre eles e desbloquear recursos, incluindo previsão, geração e raciocínio. Quanto mais rápido os tokens puderem ser processados, mais rápido os modelos poderão aprender e responder”.

diversificadas como livros, artigos acadêmicos, páginas web, códigos de programação e outros materiais textuais.

Segundo a IBM, um dataset constitui uma coleção de dados utilizados para treinar algoritmos de aprendizado de máquina, dados que podem ser estruturados (geralmente catalogados em tabelas ou rótulos) ou não estruturados, estes conjuntos de dados podem conter textos, imagens, áudios, vídeos ou combinações multimodais destes elementos, organizados de forma a facilitar o processamento algorítmico. Este processo de elaboração de dataset constitui a essência do Text and Data Mining (TDM) ou mineração de dados, que tem como escopo analisar e extrair informações desses datasets, estruturados ou não, a fim de identificar padrões, tendências e conhecimentos que serão utilizados para o treinamento dos modelos de IAG.

A obtenção destes dados, insumo substancial para treinamento de IAG, ocorre através de múltiplas metodologias, tais como coleta de dados abertos (open data), parceria com detentores de direitos, licenciamento, web scraping e geração sintética de dados. A web scraping representa a técnica mais controversa do ponto de vista jurídico, segundo Ryan Mitchell (2018), a técnica consiste na raspagem/extração automatizada de conteúdo disponível publicamente na internet através de programas computacionais que navegam sistematicamente por websites, coletando textos, imagens e outros materiais. Esta prática resulta na criação dos chamados “data scraping” que são informações extraídas sem consentimento específico dos titulares de direitos autorais.

Importa ressaltar que, no âmbito brasileiro, apesar dos conteúdos coletados via web scraping estarem publicamente disponíveis na internet, essa circunstância não implica a sua livre utilização para fins de treinamento de modelos de IAG. A Lei de Direitos Autorais, Lei n. 9.610/98 (LDA) protege as obras independentemente de registro ou forma de publicação, e exige autorização prévia e expressa do titular para qualquer ato de reprodução, mesmo parcial ou automatizada. Assim, o simples fato de uma obra estar acessível em ambiente digital não a retira da esfera de proteção jurídica, de modo que a prática de web scraping de conteúdo protegido, ainda que proveniente de sites públicos, pode configurar ato ilícito à luz da LDA.

Empresas como Common Crawl, empresa de dataset para IAG, disponibilizam datasets massivos obtidos através de web scraping, conforme informações do próprio website oficial a empresa, contendo petabytes de dados textuais extraídos de bilhões de páginas web, sendo inclusive esta a base de dados de treinamento de IAG mais utilizada no mercado, havendo informação em relatório da OCDE (2025) de que mais de 80% dos tokens utilizados para treinamento do modelo GPT-3 da OpenAI vieram da Common Crawl.

Uma das etapas mais relevantes na constituição do dataset é curadoria dos dados, que envolve processo técnico e editorial crítico que consiste na seleção, limpeza, filtragem e organização dos dados brutos coletados (embora ocorra processo de curadoria de dados em outros momentos pós-treinamento da IAG), conforme Mitchell (2019). Esta etapa visa a remoção de conteúdo duplicado, filtragem de material inadequado ou de baixa qualidade, correção de erros, balanceamento da base (para evitar vieses), anonimização e estruturação/rotulagem dos dados em formatos adequados para processamento algorítmico.

O processo de curadoria é fundamental para garantia de qualidade, confiabilidade e usabilidade dos dados, sendo motivo pelo qual há empresas especializadas na constituição e comercialização de datasets especializados.

Nesta etapa do processamento dos dados coletados, há curadoria especializada de grande massa de dados, que permite o treinamento da IAG ao objetivo que ela se destina. Este processo, se realizado adequadamente, é o que poderia garantir, por exemplo, retirada de conteúdos indevidos e prejudiciais a aprendizagem como conteúdos preconceituosos ou falsos, além de ser a técnica adequada para evitar violação de direitos autorais e privacidade.

A curadoria influencia diretamente a qualidade do aprendizado, e portanto, esta relacionada diretamente ao output ao usuário da tecnologia, assim como aos vieses do sistema e a capacidade resultante do modelo, constituindo elemento fundamental da arquitetura de IAG, pois sem tal curadoria não haveria como a IAG realizar a tarefa pretendida.

É neste aspecto do treinamento do modelo de IAG que encontra-se a intersecção com a propriedade intelectual, pois o direito autoral protege obras literárias, artísticas e científicas que apresentem originalidade e expressem a personalidade de seus autores nos termos das leis de direitos autorais, assim, no momento em que datasets incorporam textos de livros, artigos jornalísticos, fotografias, obras musicais ou outras criações protegidas, os direitos autorais originários permanecem com seus respectivos titulares, os quais se veem violados nos termos da LDA, afinal não autorizaram o uso de tais obras e tampouco perceberam qualquer pagamento por isso.

A utilização destas obras no treinamento de IAG levanta questões fundamentais sobre os limites dos direitos autorais, sejam as obras obtidas por meio de web scraping de páginas com autorização da veiculação da obra ou páginas web com conteúdo considerado pirata³, a mera

³ A pirataria, no contexto digital, consiste na reprodução, distribuição ou uso de obras intelectuais protegidas por direitos autorais sem a devida autorização dos titulares, como por exemplo a disponibilização *online* de uma cópia digitalizada de um livro físico. É uma violação da propriedade intelectual que abrange desde a cópia e distribuição

incorporação de obras protegidas em datasets pode, portanto, constituir reprodução não autorizada.

Há risco de que o processo de treinamento de IAG possa caracterizar criação de obra derivada baseada no conjunto de obras originais incorporadas ao dataset, afinal os modelos de LLM resultantes seriam aptos a gerar conteúdo que incorpora elementos, estilos ou características das obras utilizadas em seu treinamento.

Em geral, as empresas de IAG tratam seus datasets como informações confidenciais protegidas por segredo comercial em razão da importância dessa informação dentro da arquitetura da IAG, o que permite proteção de metodologias de curadoria, fontes de dados exclusivas, técnicas de processamento e outros elementos que conferem vantagem competitiva para a empresa exploradora da tecnologia.

O segredo comercial oferece proteção indefinida, empresas como OpenAI e Anthropic (proprietária do Claude) mantêm sigilo sobre composição específica, fontes e metodologias de seus datasets, invocando proteção por segredo industrial, o que dificulta significativamente compreender quais obras foram utilizadas para treinamento do modelo, impondo ônus da prova perverso aos autores das obras, detentores do direito autoral.

Estudos realizados pelo Centre for Research on Foundation Models da Stanford University (2025) identificou que somente 1 em cada 14 desenvolvedores de IAG divulgam claramente status de direitos autorais e licenciamento dos dados utilizados no treinamento de seus modelos.

O relatório da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) publicado em 2025 identifica tensões crescentes entre desenvolvimento de IAG e proteção adequada dos direitos dos autores originais dado aumento significativo de denúncias de autores que alegam ter tido suas obras utilizadas indevidamente no treinamento de modelos.

A utilização massiva de obras protegidas em datasets de treinamento levanta questões sobre remuneração apropriada para autores cujas criações contribuem para o desenvolvimento de sistemas comerciais de IAG.

Os riscos de apropriação intelectual manifestam-se através de múltiplas dimensões. Primeiramente, a incorporação não autorizada de obras protegidas em datasets comerciais pode constituir violação direta de direitos autorais. Adicionalmente, modelos de IAG treinados com obras protegidas podem reproduzir elementos substanciais destas criações, gerando preocupações

de software, música, filmes e livros até a utilização de obras para treinar modelos de IAG.

sobre competição desleal com autores originais.

ANÁLISE COMPARADA DE DIREITO AUTORAL NO USO DE OBRAS PROTEGIDAS EM DATASET DE IA

A discussão sobre o uso de obras protegidas no treinamento de inteligência artificial não pode ser compreendida sem referência ao cenário internacional. A natureza transnacional dos datasets, frequentemente obtidos por técnicas de web scraping em ambientes digitais sem fronteiras, torna impossível analisar a questão apenas sob a ótica de um ordenamento jurídico nacional. Assim, a comparação entre diferentes regimes normativos revela não apenas divergências de abordagem, mas também pontos de interseção que podem servir de guia para soluções híbridas.

O direito comparado mostra-se especialmente relevante diante da assimetria entre países de origem das principais empresas de tecnologia - Estados Unidos, Japão, China e Cingapura - e jurisdições de recepção, como o Brasil, que se deparam com a circulação global de produtos baseados em IA. Enquanto a prática de coleta e reutilização de dados se dá em escala planetária, os instrumentos jurídicos ainda permanecem fragmentados, gerando insegurança jurídica tanto para titulares de direitos quanto para desenvolvedores.

Nesse contexto, o foco recai sobre o direito autoral como eixo central da análise. O problema do uso de datasets não é, em essência, de natureza técnica, mas de compatibilidade entre os modos de aprendizado algorítmico e os limites da proteção conferida a obras intelectuais. A questão central é verificar até que ponto a incorporação de criações protegidas em bases de treinamento pode ser legitimada por exceções legais, licenciamento voluntário ou regimes de compensação obrigatória.

Além disso, a comparação revela diferentes prioridades: sistemas jurídicos mais inclinados à promoção da inovação tendem a flexibilizar a aplicação do direito autoral, enquanto regimes mais protetivos enfatizam a autonomia dos autores e o direito de controlar o uso de suas obras. Essa tensão entre fomento tecnológico e tutela da criatividade humana está no centro da disputa regulatória que hoje se desenrola nas principais jurisdições, como União Europeia que autoriza a mineração de dados (TDM) em situações específicas, enquanto Estados Unidos possuem cláusula aberta (fair use) conforme veremos adiante.

O exame comparativo não se limita a identificar normas, contribuindo para refletir sobre

a possibilidade de construção de mecanismos normativos que conciliem inovação, remuneração justa e segurança jurídica no treinamento de modelos de IAG.

Brasil

Panorama de Direito Autoral Brasileiro e Limitações ao uso Livre

O direito autoral brasileiro possui status constitucional, sendo expressamente consagrado na Constituição Federal de 1988 (CF) como direito fundamental de propriedade intelectual, no art. 5, XXVII e XXVIII, o qual constitui garantia fundamental com natureza de direito de propriedade especial, protegendo direitos individuais fundamentais como garantia do autor como titular exclusivo de sua criação, tutela aspectos morais como a criação, sem prejuízo, por óbvio, da proteção a livre iniciativa e exploração econômica do trabalho humano intelectualmente realizado.

Da mesma forma, o direito autoral é regido, pela Lei nº 9.610/1998 (Lei de Direitos Autorais - LDA), que possui inspiração no modelo europeu⁴, enfatizando na proteção da personalidade e da criação intelectual do autor, cuja legislação expressou de forma clara sua estrutura dualista, composta pelos direitos morais, de caráter inalienável e perpétuo (arts. 24 a 27 da LDA), e pelos direitos patrimoniais, que conferem ao titular o poder de explorar economicamente a obra (arts. 28 a 31 da LDA).

A natureza jurídica do direito autoral transcende a mera dimensão econômica, refletindo um vínculo indissolúvel entre o autor/criador e sua criação intelectual, o que assegura o controle sobre o uso e a difusão da obra, além do reconhecimento moral sobre o fruto de sua expressão.

A doutrina de apresentada na obra de Lincoln Antonio de Castro (2001) concebe os direitos morais do autor como modalidade de direito de personalidade, representando o aspecto pessoal do autor em relação à sua criação, ou seja, o direito ou prerrogativa de defender a obra como atributo de sua própria personalidade, uma vez que ela é a emanação de sua mais íntima

⁴ O modelo europeu denominado “Droit D’Auteur”, desenvolvido a partir de ideias filosóficas e jurídicas na França do século XVIII, como Jean-Jacques Rousseau e Denis Diderot, nesta doutrina o direito autoral é visto como um direito pessoal do criador, ligado à sua personalidade, entendendo a expressão da obra como extensão do autor. Ela possui ênfase aos direitos morais do autor, possuindo características como inalienabilidade, imprescritibilidade e irrenunciabilidade. São expressos direitos como (i) direito a paternidade - ser reconhecido como criador da obra; (ii) direito a integridade - impedir que a obra seja alterada ou distorcida de forma prejudicial a reputação do autor; (iii) direito de inédito - decidir quando e como a obra será publicada pela primeira vez; (iv) direito de retirada - permitindo a retirada de circulação da obra, ainda que já publicada; e (v) direitos patrimoniais - exploração econômica da obra.

divagação e pensamento manifestado. Essa concepção distancia o direito autoral brasileiro de uma lógica estritamente econômica passando a transcender mero aspecto mercantil do direito.

O direito autoral brasileiro compreende que as criações humanas, expressas por qualquer meio e fixadas em qualquer suporte (tangível ou intangível), sejam obras literárias, artísticas ou científicas são protegidas pela Lei de Direito Autoral, tais obras denominadas pela doutrina como “criações do espírito” são objeto de proteção legal independentemente de registro ou qualquer ato formal, bastando que tal criação se externalize ao mundo de forma fixa, ou seja, não transitória.

O prazo de proteção dos direitos patrimoniais decorrentes da obra é temporário, com prazo de duração durante toda vida do autor mais 70 anos, contados a partir do dia 01 de janeiro do ano subsequente a morte do autor nos termos do art. 41 da LDA. Exceto obras audiovisuais e fotográficas cujo prazo de 70 anos passa a ser contado de 1º de janeiro do ano subsequente ao de sua divulgação (art. 44 da LDA).

Após o transcurso do prazo de vigência dos direitos patrimoniais, a obra pertencerá a domínio público, assim como obras de autores falecidos que não tenham deixado herdeiros e obras de autores desconhecidos (denominadas “obras órfãs”) conforme previsão expressa na legislação.

Portanto, para uso de obras protegidas é necessária prévia e expressa autorização, ou licença, concedida pelo autor ou de seus herdeiros, salvo nas hipóteses legais previstas taxativamente no art. 46 da LDA, tais como usos para fins de estudo/critica, jornalísticos e cópia de trechos para uso privado (sem intuito de lucro), o que restringe significativamente a flexibilidade interpretativa, não havendo em que se falar em interpretação extensiva da norma ou aplicação de uma doutrina aberta do fair use - cláusula que permite usos não autorizados de obras quando considerados transformativos ou socialmente relevantes -, o ordenamento brasileiro não reconhece exceções genéricas de uso justo.

No Brasil não há doutrina do fair use, como ocorre no ordenamento americano, sendo imprescindível observância estrita da legislação nas hipóteses de uso livre das obras protegidas.

Não obstante, é evidente que os direitos autorais previstos na legislação e expressos constitucionalmente submetem-se a um binômio ou paradoxo, pois ao passo que tem por objetivo tutelar os direitos do autor (sejam eles, morais ou patrimoniais), de outro lado também precisam garantir que a sociedade tenha acesso a essas obras, fomentando a cultura e o conhecimento.

Assim, as limitações aos direitos do autor expressos na LDA devem ser entendidas como

confronto entre interesses de fundo constitucional que exigem ponderação e racionalidade (e não apenas aplicação restritiva da norma pura e simples), conforme podemos verificar em decisão proferida pelo Egrégio Superior Tribunal de Justiça:

“[...]O direito de exclusividade é, entretanto, intrinsecamente limitado pela sua função social, de modo que o âmbito de proteção efetiva do direito à propriedade autoral ressaia após a consideração das limitações contidas nos arts. 46, 47 e 48 da Lei 9.610/1998, interpretadas e aplicadas de acordo com os direitos e garantias fundamentais. Por esse motivo, valores como a cultura, a ciência, a intimidade, a privacidade, a família, o desenvolvimento nacional, a liberdade de imprensa, de religião e de culto devem ser considerados quando da conformação do direito à propriedade autoral. O objetivo último do direito de autor não é assegurar a riqueza material do titular do direito, mas, sim, promover a criatividade intelectual e o enriquecimento cultural da sociedade, de modo que, ainda sob essa razão, a exceção ao uso exclusivo dos direitos de autor deve se limitar ao que é estritamente necessário ao interesse público.[...]” (REsp n. 2.143.010/SP, relator Ministro Ricardo Villas Bôas Cueva, Terceira Turma, julgado em 10/6/2025, DJEN de 23/6/2025.)

Nesse contexto, deve-se observar a Regra dos Três Passos, prevista na Convenção de Berna (art. 9.2), da qual o Brasil é signatário (Decreto nº 75.699/1975), e incorporada ao Acordo TRIPS (art. 13, ratificado pelo Decreto nº 1.355/1994). Essa regra permite limitações aos direitos exclusivos do autor apenas em casos especiais (passo 1), que não conflitem com a exploração normal da obra (passo 2) e não prejudiquem injustificadamente os interesses legítimos do titular (passo 3). Embora o Brasil não adote o fair use, o Judiciário aplica essa regra para uma interpretação sistemática da LDA, equilibrando direitos fundamentais.

Conforme leciona Denis Barbosa (2012), embora o Brasil não tenha acolhido a doutrina do fair use, a aplicação da Regra dos Três Passos pelo Judiciário busca uma interpretação sistemática e teleológica da LDA, observadas as limitações estabelecidas na norma e garantindo a tutela dos direitos fundamentais em colisão com o direito do autor.

Assim, no contexto de IAG, o treinamento de modelos com obras protegidas - ainda que para fins de deep learning - demandam a necessidade de autorização prévia e expressa do titular, sob pena de configurar violação autoral.

A ausência de um dispositivo análogo ao fair use implica que o uso de obras protegidas em datasets de IAG é, em regra, ilícito sem consentimento, independentemente de o processo ser automatizado (via web scraping ou não). Essa rigidez contrasta com o sistema norte-americano, que adota a doutrina do fair use (17 U.S.C. § 107), permitindo usos não autorizados quando transformativos ou de utilidade social, o que pode abranger o treinamento de modelos de IAG, embora haja debates em curso na jurisprudência e doutrina, com posicionamentos variados sobre

se tal uso constitui fair use ou requer licenciamento.

Importa salientar que a prática de web scraping com captura e uso de conteúdo protegido pela LDA pode, em determinados casos, ser objeto de configuração de ilícito penal, nos termos do art. 184 do Código Penal (CP) com pena de detenção de 3 meses a 1 ano, e multa, denominado contrafação. A contrafação, também prevista no artigo 5º, VII da LDA, configura-se como a reprodução não autorizada de obra intelectual, integral ou parcial, com intuito de obtenção de vantagem direta ou indireta.

No ambiente digital, essa prática assume novas formas, que vão desde os sites piratas voltados à distribuição massiva de livros, músicas e filmes até mecanismos automatizados de web scraping que copiam e armazenam, sem consentimento, conteúdos protegidos para compor bases de dados de modelos de IAG.

Ainda que parte dessas obras estejam publicamente acessível na internet, a mera disponibilidade não as torna livres de proteção autoral, pois o direito do autor à autorização prévia permanece íntegro. O uso de obras hospedadas em portais de notícias, repositórios acadêmicos ou plataformas culturais sem licenciamento constitui, portanto, hipótese de contrafação digital, sujeita às sanções civis e penais cabíveis, inclusive a previstas no art. 184 do CP, que criminaliza a violação de direito autoral por meio de reprodução não autorizada.

Além dos casos clássicos de pirataria, a contrafação assume contornos mais sutis no contexto da IAG, em que a web scraping sistemático de obras brasileiras disponíveis na web, ainda que em sites legítimos, implica reprodução técnica e armazenamento duradouro em bases de dados proprietárias. Essa conduta, à luz da LDA e da Regra dos Três Passos (da Convenção de Berna), extrapola as limitações legítimas de uso, pois interfere na exploração econômica normal da obra e prejudica os interesses legítimos do autor.

Assim, o treinamento de modelos de IAG com conteúdo extraídos da internet sem autorização poderia vir a ser configurada prática de contrafação digital, afrontando os direitos patrimoniais do autor e o equilíbrio normativo consagrado pela lei brasileira e tratados internacionais.

Além disso, entidades de propriedade intelectual como o U.S Copyrights Office (2025) não descartam a possibilidade do fenômeno da diluição de mercado autoral, cujo uso massivo de obras humanas para treinar modelos capazes de reproduzir estilos ou formatos reduz o valor de mercado dos autores originais, comprometendo a sustentabilidade da produção cultural.

Assim, o direito autoral brasileiro, baseado em uma concepção personalista e

patrimonialista restritiva, não comporta, sob o regime atual, o uso de obras protegidas em datasets de IAG sem autorização expressa. A rigidez da LDA, aliada à ausência de exceções voltadas à pesquisa e à inovação, coloca o Brasil em posição de desvantagem regulatória quando comparado a sistemas que já incorporaram salvaguardas tecnológicas específicas, essa constatação explica a crescente pressão por atualização legislativa, refletida nos debates em torno do Projeto de Lei nº 2.338/2023, que propõe um marco legal para a inteligência artificial, ainda que de forma incipiente no tocante ao direito autoral, conforme veremos a diante.

Portanto, a ausência de exceções tecnológicas específicas na LDA evidencia a urgência de reforma normativa que harmonize o incentivo à inovação digital com a tutela dos direitos autorais, sob pena de o país permanecer à margem do debate global sobre governança da inteligência artificial.

Caso OpenAI e Folha de S. Paulo

O processo judicial movido pela Empresa Folha da Manhã S.A., editora do jornal Folha de S. Paulo, contra as empresas OpenAI, Inc., OpenAI Global, LLC, e OpenAI GP, LLC representa o primeiro grande “teste” da legislação autoral brasileira diante da ascensão da inteligência artificial generativa. Trata-se de um litígio paradigmático, capaz de definir como o direito autoral nacional responderá à incorporação de obras jornalísticas e literárias em modelos de aprendizado de máquina, especialmente num país que, diferentemente dos Estados Unidos, não adota a doutrina aberta do fair use.

A ação coloca frente a frente dois valores constitucionais igualmente relevantes: de um lado, a proteção da criação intelectual e da livre iniciativa; de outro, a promoção da inovação e do acesso à informação. Por seu ineditismo, o caso tende a consolidar o primeiro precedente brasileiro sobre o uso de conteúdos protegidos em datasets de sistemas de IAG.

A demanda foi proposta pela Folha de S. Paulo, em agosto de 2025, sob o número 1107237-96.2025.8.26.0100, perante a 3ª Vara Empresarial e de Conflitos de Arbitragem de São Paulo. Na petição inicial, a autora sustenta que as rés teriam utilizado, sem autorização e sem remuneração, grande volume de reportagens e textos publicados no portal folha.uol.com.br para o treinamento de seus modelos de linguagem, o ChatGPT, incluindo entre os alegados conteúdos utilizados aqueles reservados exclusivamente para assinantes da Folha de S. Paulo.

Afirma-se que a reprodução e o uso dessas obras configuram violação aos direitos

patrimoniais e morais previstos nos artigos 24, 28 e 29 da LDA, e prática de concorrência desleal nos termos do artigo 195, inciso III, da Lei da Propriedade Industrial (LPI). Argumenta ainda que a apropriação indevida de seu acervo jornalístico interfere diretamente em sua exploração econômica, uma vez que os sistemas das rés passam a oferecer respostas e resumos baseados nas mesmas informações, desviando audiência e reduzindo o valor de mercado do produto jornalístico.

A Folha requer a imediata cessação da coleta e do uso de seu conteúdo, a exclusão dos dados de seus repositórios, indenização por danos materiais e morais, e a destruição dos modelos de IAG construídos a partir do treinamento feito com base no conteúdo da autora.

Em 03/09/2025, o juiz Fábio Henrique Prado de Toledo indeferiu o pedido de tutela de urgência e determinou o segredo de justiça sobre os autos. A decisão reconheceu que a controvérsia envolve “questão de alta complexidade” e matéria fática e jurídica ainda não analisada nos tribunais brasileiros, especialmente no tocante à identificação das fontes de dados e à estrutura interna dos sistemas de IA, cuja divulgação poderia expor segredos comerciais e informações técnicas sensíveis. Assim, com base nos artigos 189, inciso IV, do Código de Processo Civil, e 206 da LPI, o magistrado deferiu o sigilo, ponderando a oposição da autora ao segredo de justiça, que alegou a regra geral da publicidade dos atos processuais e ausência de segredo de indústria.

Quanto à liminar, entendeu não demonstrado o perigo de dano irreparável, por se tratar de questão de natureza econômica passível de reparação posterior. A decisão também sinalizou que qualquer medida restritiva sobre sistemas de demandaria base probatória robusta, diante da ausência de precedentes e da complexidade técnica do tema, e designou audiência de conciliação.

Em manifestação, a OpenAI alega a licitude de suas práticas, sustentando que o treinamento de modelos não constitui reprodução de conteúdo protegido, mas sim uma operação técnica de “aprendizado estatístico” sobre padrões linguísticos. Alegou que o processo de tokenização, etapa que transforma palavras e frases em unidades matemáticas abstratas, não implica armazenamento permanente ou leitura das obras originais, mas apenas uso de fragmentos com finalidade de aprendizado.

Sob a ótica abordada pela OpenAI, o treinamento seria uma forma de uso transformativo, amparada pela Regra dos Três Passos prevista na Convenção de Berna (art. 9.2) e no Acordo TRIPS (art. 13), já que não substituiria a função econômica da obra nem prejudicaria seus interesses legítimos. A empresa também argumentou que a coleta de dados ocorreu em

conformidade com padrões internacionais de acesso público à informação, alega também que a autora implementou mecanismos técnicos de bloqueio (robots.txt) para impedir a coleta de dados, e que o portal da Folha não havia bloqueado integralmente o rastreamento automatizado à época dos fatos.

O caso também evidencia a delicada questão da territorialidade no uso de obras brasileiras por empresas estrangeiras cujos modelos de são desenvolvidos e hospedados fora do país. Embora a coleta tenha ocorrido em sites com conteúdo em português realizados por autores sob égide da LDA, mas o processamento dos dados e o treinamento dos modelos ocorrem em servidores situados em outros países, o que suscita dúvidas sobre o alcance da LDA e sobre a jurisdição competente para julgar eventuais violações. Ainda assim, o fato de o conteúdo ter origem e destinação voltadas ao público brasileiro, considerando o conteúdo e idioma do output, reforça a aplicação da legislação nacional e a legitimidade da atuação do Judiciário local.

Ademais, a prática de web scraping em larga escala, realizada sem consentimento, permanece no centro da controvérsia: para a autora, ela constitui contrafação (crime previsto no art. 184 do Código Penal) e concorrência desleal; para as rés, trata-se de coleta técnica de dados abertos, necessária ao desenvolvimento de tecnologia baseada em linguagem natural.

Desta forma, com ângulos diametralmente opostos, embora compreensíveis de parte a parte, sob aspecto sociotecnológico jurídico, a disputa expõe o ponto de fricção entre o conceito clássico de “reprodução”, central na LDA, e o funcionamento dos modelos de IA, que não armazenam as obras em sua integridade, mas as fragmentam e processam em tokens matemáticos, cujo procedimento, ainda que não reproduza textos identificáveis, depende de cópias temporárias durante o aprendizado, gerando debate sobre se tais cópias configuram ou não uso ilícito.

A importância do caso Folha de S. Paulo v. OpenAI transcende a relação entre as partes, afinal tal processo inaugura no Brasil o debate judicial sobre o limite do uso de dados autorais em sistemas de IAG, a responsabilidade das plataformas pelo uso de obras protegidas em seus modelos e a necessidade de transparência nas fontes de treinamento. Ao mesmo tempo, evidencia os desafios da aplicação do direito autoral em escala global, em um cenário em que a coleta e o processamento de informações não se submetem facilmente a fronteiras territoriais.

Uma decisão favorável a Folha de S. Paulo, configuraria uma vitória aos autores, e poderá ser precedente relevante para que empresas de obtenham autorizações prévias ou celebrem acordos de licenciamento para utilizar conteúdo brasileiro, encarecendo o desenvolvimento de modelos e potencialmente limitando sua aplicação no país.

Por outro lado, uma decisão favorável às ré, OpenAI, poderia ser interpretado por empresas de tecnologia como sinalização judicial de que o treinamento de modelos não constitui violação autoral, encorajando investimentos em IAG, mas potencialmente desprotegendo autores.

É neste paradoxo que será proferida decisão deste processo emblemático, afinal a decisão judicial transitada em julgado, qualquer que seja, poderá ser utilizada como *leading case* para demonstrar os limites e, eventualmente, o equilíbrio entre proteção intelectual e liberdade tecnológica, influenciando tanto a elaboração de políticas públicas quanto o desenho de futuras reformas legislativas sobre IAG no Brasil.

Debates sobre o PL nº 2.338/2023 e o Direito Autoral na IAG

Enquanto o cenário mundial ainda se delineia com a franca expansão da IAG sendo introduzida no cotidiano da sociedade, o Brasil posiciona-se entre os países de vanguarda que buscam legislar a temática com o chamado Marco Legal da Inteligência Artificial, o Projeto de Lei nº 2.338/2023 (PL), aprovado no Senado Federal em 2024, e atualmente em tramitação na Câmara dos Deputados, cujo enfoque está voltado para princípios éticos, segurança e responsabilidade.

Embora o texto do PL não seja destinado à regulação da propriedade intelectual, seu conteúdo repercute de modo direto nas discussões sobre direito autoral e utilização de obras protegidas no treinamento de modelos de IAG. O texto atual, conforme a versão consolidada de 17/03/2025, dedica o Capítulo IX (arts. 62 a 66) tratam da proteção dos direitos autorais e do uso de obras intelectuais em contextos de IA apresentando diretrizes gerais que visam conciliar o avanço tecnológico com a preservação dos direitos de autores.

O artigo 62 do PL reafirma a aplicabilidade das normas de direito autoral, estabelecendo que o uso de obras protegidas por sistemas de deve observar a LDA, Convenção de Berna e o Acordo TRIPS.

O PL prevê que a utilização de obras intelectuais em processos de treinamento, validação ou operação de sistemas de depende de autorização prévia dos titulares de direitos ou de entidades de gestão coletiva, salvo nas hipóteses expressamente previstas em lei, em geral voltadas para tecnologias sem finalidade lucrativa, como P&D por exemplo.

Introduz um dever de transparência, impondo que os desenvolvedores mantenham registros documentais sobre a origem e o licenciamento das obras utilizadas em datasets de

treinamento, obrigação que se harmoniza com os princípios de rastreabilidade e accountability constantes do Capítulo III do mesmo projeto.

O PL prevê a possibilidade do Poder Executivo, mediante regulamentação, criar mecanismos de compensação coletiva para titulares de direitos autorais cujas obras sejam utilizadas no treinamento de sistemas de IA, abrindo espaço para a constituição de um modelo de licenciamento coletivo digital, cuja proposta remete à experiência brasileira com o ECAD e aos mecanismos de gestão coletiva existentes na União Europeia, é apontada por autores contemporâneos como alternativa viável para equilibrar a proteção do autor e a viabilidade econômica das empresas de IA.

O artigo 66 do PL encerra o capítulo determinando que o uso não autorizado de obras protegidas nos processos de treinamento ou operação de constitui violação aos direitos autorais e sujeita o infrator às sanções civis e criminais previstas na legislação vigente.

Não obstante o texto do Projeto, a efetividade de tais dispositivos dependerá da regulamentação posterior a fim de operacionalizar a compensação coletiva e definir parâmetros técnicos para o registro e rastreabilidade das fontes utilizadas no treinamento de modelos de IAG. A ausência de mecanismos verificáveis de auditoria e de transparência sobre a composição dos datasets, em sua maioria das vezes tratados como segredos comerciais, pode transformar a obrigação legal de informação em mero enunciado formal, perpetuando o fenômeno das “caixas-pretas algorítmicas” e dificultando a fiscalização do cumprimento dos direitos autorais em ambiente digital.

Assim, os debates em torno do PL evidenciam que o projeto e a sociedade ainda carecem de detalhamento técnico e jurídico sobre aspectos específicos do direito autoral aplicado à IAG. Os debates legislativos realizados através das audiências públicas realizadas ano longo de 2025 na Comissão Especial sobre Inteligência Artificial no Câmara Legislativa, revelam a tentativa de conciliar a preservação da autoria humana com a dinâmica de aprendizado automatizado dos modelos generativos, mas a ausência de clareza sobre o funcionamento do treinamento da IAG e parâmetros específicos sobre o que constitui uso lícito ou ilícito de obras protegidas em datasets gera incerteza regulatória e pode transferir ao Judiciário o papel de interpretar caso a caso a compatibilidade entre inovação tecnológica e direito autoral.

O PL nº 2.338/2023, portanto, marca um passo importante, embora ainda preliminar, na adequação do ordenamento jurídico brasileiro à realidade da inteligência artificial. Ao incluir dispositivos específicos sobre o uso de obras protegidas, o legislador reconhece que a não é neutra

quanto à propriedade intelectual e que o treinamento de modelos depende de material humano preexistente.

Se aprovado o Projeto de Lei, a efetividade das normas dependerá da regulamentação posterior, especialmente quanto aos mecanismos de compensação, transparência e fiscalização. A consolidação desses instrumentos poderá definir se o Brasil adotará um modelo protetivo de inspiração continental, centrado no consentimento do autor, ou se evoluirá para um regime mais flexível e funcional, inspirado em princípios de uso legítimo e licenciamento coletivo, em ambos os casos, a tensão entre inovação e tutela autoral permanecerá como eixo de equilíbrio do futuro regime jurídico da IA.

Estados Unidos (EUA)

Panorama de Copyright no EUA

O direito autoral nos Estados Unidos, conhecido como copyright, representa um dos pilares fundamentais da propriedade intelectual, fundamentado no Artigo I, Seção 8, da Constituição Americana, que confere ao Congresso o poder de promover o progresso das ciências e artes úteis por meio da concessão de direitos exclusivos limitados aos autores. Essa proteção é codificada principalmente no Copyright Act de 1976, que estabelece o quadro normativo atual, substituindo legislações anteriores e alinhando-se a tratados internacionais como a Convenção de Berna.

O copyright visa equilibrar os interesses econômicos dos autores, incentivando a produção cultural e intelectual, com o bem público, permitindo o acesso e a disseminação de conhecimentos. Diferentemente do modelo brasileiro e sistemas europeus, que enfatizam direitos como o direito à integridade da obra, o modelo americano prioriza aspectos utilitários e econômicos, concedendo aos titulares o controle sobre a reprodução, distribuição, performance pública, exibição e criação de obras derivadas de suas criações originais.

Os conceitos centrais do copyright giram em torno da proteção automática de "obras originais de autoria" fixadas em um meio tangível, abrangendo categorias como obras literárias, musicais, dramáticas, pictóricas, gráficas, escultóricas, cinematográficas, sonoras e arquitetônicas, conforme delineado na Seção 102(a) do Copyright Act. A proteção surge no momento da fixação - por exemplo, ao escrever um texto em papel ou salvar um arquivo digital

-, sem necessidade de registro formal junto ao United States Copyright Office (USCO), embora o registro seja aconselhável para fins probatórios e acesso a medidas judiciais, como danos estatutários de até US\$ 150.000 por infração intencional (Seção 504).

Os requisitos para a proteção autoral incluem, primordialmente, a originalidade, que demanda um mínimo de criatividade independente, sem exigência de novidade absoluta ou mérito artístico, como esclarecido pela Suprema Corte no precedente *Feist Publications, Inc. v. Rural Telephone Service Co.* (1991), onde se negou proteção a listas telefônicas por ausência de seleção criativa.

Outro requisito implícito é a autoria humana, excluindo obras geradas puramente por inteligência artificial, conforme *Compendium of U.S. Copyright Office Practices* (2021), no qual afirma que o copyright protege apenas obras de criação humana.

O prazo de proteção varia conforme a data de criação, para obras pós-1º de janeiro de 1978, estende-se pela vida do autor mais 70 anos, já para obras anônimas, pseudônimas ou feitas sob encomenda o prazo é de 95 anos da publicação ou 120 anos da criação, o que for menor, refletindo ajustes do Copyright Term Extension Act de 1998.

Já o domínio público constitui o contraponto ao copyright, abrangendo materiais não protegidos por leis de propriedade intelectual, permitindo uso livre, reprodução, adaptação e distribuição sem permissão ou compensação ao autor original. Conceitualmente, o domínio público enriquece o patrimônio cultural coletivo, servindo como reservatório de ideias e expressões que fomentam novas criações, alinhando-se ao objetivo constitucional do copyright de promover o progresso social.

Fatores Preponderantes para Caracterização do Fair Use

Há hipóteses em que, pela doutrina americana, é possível fazer uso de obras protegidas por copyright sem autorização ou compensação ao autor da obra, o qual denomina-se fair use, em tradução literal, uso justo.

A doutrina do fair use é expressa na Seção 107 do Copyright Act, ela representa uma limitação essencial aos direitos exclusivos do autor, permitindo uso de obras protegidas, dentre outros usos, para fins como crítica, comentário, reportagem, ensino, bolsa de estudos ou pesquisa, promovendo o equilíbrio entre proteção autoral e liberdade de expressão.

Sua aplicação é flexível e casuística, avaliada judicialmente por meio de quatro fatores

principais, os quais são expressos na seção do Copyright Act, sendo tais fatores (1) o propósito e caráter do uso, incluindo se é comercial ou não lucrativo/educacional, e se adiciona novo significado ou propósito; (2) a natureza da obra protegida, favorecendo usos de materiais factuais em comparação com publicações de obras fictícias ou inéditas; (3) a quantidade e substancialidade da porção utilizada da obra protegida, considerando não só o volume quantitativo, mas a importância qualitativa, como o "coração" da obra; e (4) o efeito sobre o mercado potencial ou valor da obra original, analisando se o uso substitui, prejudica ou estimula o mercado econômico em que esta inserido a obra protegida.

Um aspecto importante nos precedentes modernos na doutrina do fair use, é o caráter transformativo do conteúdo, enfatizado pela Suprema Corte em precedentes como *Campbell v. Acuff-Rose Music, Inc.* (1994), em que se estabeleceu que um uso é transformativo quando adiciona algo novo, com propósito ou caráter diferente, alterando a obra original com nova expressão, significado ou mensagem, mesmo em contextos comerciais como paródias.

Os tribunais americanos avaliam de forma conjunta os cinco fatores relacionados à admissibilidade do fair use. A presença isolada de apenas um dos fatores ou parte deles não é suficiente para caracterização do uso justo, requerendo que seja realizada uma análise holística para determinar a aplicabilidade da doutrina.

É manifesta a relevância do caráter transformador quando se trata de tecnologia que deverá ser avaliado em conjunto com os demais fatores, vejamos o que ocorreu no caso *Google LLC v. Oracle America, Inc.* (2021), em que a Suprema Corte considerou transformativo o reuso de trechos copiados do código de aplicação ("API") Java pela uso da Google no então criado Android. Esta decisão se deu por entender que esses trechos copiados foram recontextualizados em um novo ambiente funcional, com finalidade diferente da obra original. Sendo assim, o critério não é apenas se houve cópia, mas se a obra copiada foi incorporada a um novo propósito, expressão ou utilidade, sem substituir a original.

O fator transformativo, portanto, funciona como um "atenuante", podendo mitigar preocupações com usos comerciais e favorecer usos inovadores, como aspectos educacionais e críticos, embora exija análise rigorosa para evitar que abusos esvaziem os direitos econômicos do autor.

Fair Use na IAG

Essa doutrina, desenvolvida pela jurisprudência da Suprema Corte em precedentes como *Campbell v. Acuff-Rose Music Inc.* (1994) e *Google LLC v. Oracle America Inc.* (2021), a Seção 107 do Copyright Act prevê expressamente a possibilidade de uso não autorizado, desde que presente determinados fatores, quais sejam, (Fator 1) finalidade e o caráter do uso, incluindo natureza comercial, (Fator 2) natureza da obra protegida, (Fator 3) quantidade e substancialidade da obra em relação a obra protegida por direitos autorais como um todo, e (Fator 4) efeito do uso sobre o mercado potencial ou valor da obra protegida por direitos autorais.

No contexto da IAG, o fair use emerge como o eixo central de controvérsias sobre a incorporação de obras protegidas em datasets de treinamento, obtidos predominantemente via web scraping ou, em menor escala, por meio de licenças negociadas, que representam exceções contratuais à regra da autorização prévia exigida pela Seção 106.

A aplicação do fair use ao treinamento de IAG é particularmente intrincada, pois o processo envolve cópias de obras protegidas para treinamento da IA, potencialmente implicando violações dos direitos de reprodução e de preparação de obras derivadas (Seção 106(2) do Copyright Act), mesmo que os outputs finais da IAG não reproduzam integral ou parcialmente a obra utilizada como material de treinamento do modelo.

Estudiosos como Edward Lee, em seu artigo *Fair Use and The Origin of IA Training* argumentam que o treinamento constitui um uso transformativo sob o primeiro fator do fair use, sendo análogo à indexação em motores de busca ou à análise de dados em pesquisas acadêmicas, pois o propósito primordial do uso da obra no aprendizado da IAG é a extração de padrões estatísticos para gerar novas saídas de texto, e não, necessariamente, reproduzir os originais, o que estaria alinhado ao objetivo constitucional do copyright de incentivar a inovação e a criação intelectual.

Contudo, críticos como Kevin Madigan (2024), argumentam que a aplicação da doutrina do fair use ao treinamento de IAG é problemática, sobretudo pelo caráter eminentemente comercial desses modelos, o que pesaria contra o primeiro fator do fair use, já que haveria exploração econômica que eventualmente poderia descaracterizar o uso justo da obra. Essa posição destaca a necessidade de maior cautela na extensão de exceções autorais a usos massivos e lucrativos, especialmente por empresas que monetizam seus modelos através de assinaturas para usuários e aplicações de uso corporativo.

Datasets utilizados nos modelos de IAG englobam bilhões de tokens obtidos por web scraping, de fontes como Common Crawl, não havendo distinção entre conteúdo licenciado e conteúdo pirata, o que poderia potencializar o impacto no mercado da obra original, ocasionando inclusive diluição da demanda⁵.

O relatório Generative AI Training da USCO (2025), oferece um panorama técnico-jurídico denso sobre o tema, concluindo que o uso de obras protegidas por direitos autorais que compõem datasets de treinamento podem ferir os direitos exclusivos de reprodução dos autores, e ser entendido como criação de obras derivadas, o que requer uma análise casuística do fair use sem aplicação categórica ou automática, dependendo do fator transformador e impactos no mercado.

Este relatório é resultado de consultas públicas extensas e análise de submissões de stakeholders, e delineia o processo de treinamento como envolvendo tais “cópias intermediárias” - como vem sendo chamados os outputs gerados pela IAG - que, embora transitórias, satisfazem os requisitos de fixação e originalidade para proteção autoral, potencialmente violando a Seção 106 mesmo em casos de web scraping de conteúdo público.

Especificamente, o USCO enfatiza que, o fator transformador, favorece usos não comerciais ou de pesquisa, mas o caráter comercial das plataformas de IAG e a escala massiva do scraping - frequentemente sem filtros para separar e retirar materiais piratas - inclinam contra o fair use, priorizando o desenvolvimento de mercados de licenciamento voluntário e considerando soluções como licenciamento coletivo, demandando intervenção governamental imediata.

O relatório distingue licenças voluntárias, como o acordo de 2023 entre OpenAI e Associated Press para conteúdo jornalístico, como caminhos lícitos que evitam litígios, mas nota sua raridade, pois a quantidade massiva de obras usadas sem permissão é relevante, considerando o terceiro fator presente na Seção 107 do Copyrights Act, restando manifesto que a doutrina do fair use não é uma "licença gratuita" para scraping em massa de conteúdo.

Não obstante, o relatório da OCDE, publicado em fevereiro de 2025, o uso de técnicas de scraping em massa são recorrentes, culminando em diversos datasets contaminados com conteúdo

⁵ A diluição da demanda (“*Dilution Market*”), segundo US Copyright Office (2025), é uma teoria americana na qual acredita-se que o uso de obras protegidas por direitos autorais no treinamento de modelos de IA generativa pode diluir a demanda por obras originais ao fornecer substitutos competidores, mesmo que as saídas da IA não sejam substancialmente semelhantes a qualquer trabalho específico, resultando em maior competição por vendas, dificuldade para o público em identificar conteúdos humanos e redução no valor de mercado das obras, conforme analisado no fator quatro da doutrina do *fair use*.

pirata, como é o caso do dataset Books3 citado no relatório por contém mais de 170.000 livros piratas obtidos de BitTorrent não autorizados, utilizados para treinar modelos como o Llama da Meta e o GPT-J da EleutherAI.

As implicações legais e éticas da pirataria em datasets obtidos por web scraping, total ou parcial, para treinamento são profundas, gerando litígios crescentes e preocupações governamentais, uma vez que inexistente posicionamento pacífico dos tribunais americanos sobre este aspecto no contexto de aprendizado de IAG. O relatório da OCDE observa que o scraping de sites associados à pirataria pode implicar violações de direitos autorais, a banco de dados e segredos comerciais, com ações judiciais crescentes em todo mundo.

Em mapa realizado setembro de 2025 foram identificados mais de 50 processos na justiça americana envolvendo o tema de copyright em razão do uso de obras protegidas para treinamento de IA, os quais em sua maioria possuem alegação de que as obras teriam sido obtidas por meio de scraping em páginas com conteúdo pirata, conforme veremos a seguir.

Precedentes Envolvendo IAG e Aplicação do Fair Use

O panorama jurídico dos Estados Unidos relativo aos direitos autorais no contexto do treinamento de IAG, nos últimos 2 (dois) anos envolvendo diretamente inteligência artificial, tem sido profundamente influenciado pelas disputas envolvendo o uso de data scraping para a criação de datasets mistos que combinam obras licenciadas, de domínio público e, frequentemente, materiais piratas.

O caso Thomson Reuters v. ROSS Intelligence, decidido em 11 de fevereiro de 2025 pelo Tribunal Distrital de Delaware nos Estados Unidos, ilustra uma rejeição contundente do fair use em um contexto de scraping comercial. A Thomson Reuters acusou a empresa ROSS Intelligence de extrair sistematicamente headnotes (resumos de processos judiciais) protegidos por copyright do sistema Westlaw (plataforma de pesquisa jurídica da Thomson Reuters), utilizando-os para treinar um motor de busca concorrente pertencente a ROSS Intelligence.

Neste caso, o Juiz Stephanos Bibas, em sua análise dos quatro fatores do fair use, concluiu que o uso não poderia ser caracterizado como transformador (fator 1), pois replicava a funcionalidade comercial original do Westlaw, sem agregar contribuição autoral expressiva, e ainda, agravado pela natureza lucrativa do projeto realizado pela ROSS Intelligence com escopo de concorrer com o mercado da Westlaw. Ao analisar o fator 3 da doutrina do fair use, sobre a

quantidade de conteúdo usado, verificou-se que havia sido extraído milhões de páginas da Westlaw, e, quando analisado o efeito no mercado (fator 4), restou inequívoco ao juiz a diluição da demanda pelo Westlaw, comprovado pelas provas colacionadas aos autos, em razão de queda significativa no fluxo de usuários da plataforma, acarretando perda de oportunidades de licenciamento, o que culminou com o afastamento do fair use em desfavor da ROSS Intelligence, com uma indenização de US\$ 100 milhões para a Thomson Reuters e ordem de cessão imediata da conduta.

Em contraste, no caso *Kadrey v. Meta Platforms, Inc.*, julgado em 25 de junho de 2025 pelo Tribunal do Distrito Norte da Califórnia, houve apreciação dos fatores do fair use no treinamento de IAG sob perspectiva mais favorável a inovação, neste processo os autores Sarah Silverman e Richard Kadrey alegaram que a Meta utilizou livros obtidos em sites piratas para treinar o modelo LLaMA. O Juiz Vince Chhabria, ao analisar o caso entendeu pela aplicação da doutrina do fair use para o processo de treinamento, fundamentando que uso era altamente transformador (fator 1), pois visava aprendizado estatístico, e não produzia substitutos diretos a obra protegida, não obstante o caráter pecuniário da tecnologia.

Contudo, o Juiz Chhabria permitiu que a ação prosseguisse para avaliar se os outputs do modelo de IAG, poderiam ser capazes de gerar risco de diluição de mercado, na hipótese de obras derivadas, com estilo ou teor similar, possam vir a competir no mercado com as obras originais.

Outro caso relevante e inédito, é o caso *Bartz v. Anthropic PBC*, decidido em 23 de junho de 2025 pelo mesmo tribunal da Califórnia, reforça a aceitação do fair use em cenários de treinamento, mas com nuances diferenciadas. Neste caso os autores acusaram a Anthropic de usar livros piratas do dataset “The Pile” para treinar o modelo de IA, Claude.

O Juiz William Alsup considerou no caso em tela o uso transformador (fator 1), destacando que o treinamento cria modelos probabilísticos sem reprodução literal, sendo a quantidade usada (fator 3) reconhecida como necessária para a escala da IAG. Quando da análise do fator 4, o juiz rejeitou alegações de diluição de mercado, mas abriu espaço para análise de outputs gerados pelo Claude aos usuários, o que resultou, poucos dias após a prolação da decisão em acordo de US\$ 1,5 bilhão aos autores, o qual ficou conhecido como o maior acordo envolvendo copyright que se tem notícia.

A decisão do juiz Alsup enfrentou o tema da pirataria de forma não usual, afirmando que a presença de pirataria em datasets não é automaticamente fatal ao fair use, desde que o propósito seja transformador e o impacto no mercado da obra original seja mínimo.

A análise dos fatores do fair use nessas decisões revelam um padrão emergente: o Fator 1 (propósito e caráter do uso) é o pivô, com a análise do aspecto transformativo como ponto central. No caso Thomson Reuters, o uso comercial direto e a falta de inovação expressiva levaram à rejeição, alinhando-se ao precedente *Andy Warhol Foundation v. Goldsmith* (598 U.S. 508, 2023), que prioriza o mercado original em detrimento da inovação. Já no caso envolvendo a Anthropic, os juízes adotaram a lógica de *Authors Guild v. Google* (804 F.3d 202, 2015), onde o treinamento para pesquisa é visto como transformador, mesmo com datasets mistos.

Isso reflete a tensão entre a meta constitucional de promover o progresso e a proteção dos autores, com os tribunais californianos mostrando maior abertura à inovação tecnológica em comparação com outros estados, o que pode se justificar inclusive por sua posição econômica e disruptiva como “berço de startups” nos Estados Unidos.

Outro aspecto relevante a ser observado advém do fator 4, referente aos efeitos no mercado, no caso Thomson Reuters a diluição da demanda pelo Westlaw foi comprovada por evidências de perda de assinantes, enquanto no caso Bartz contra Anthropic a ausência de substituição direta das obras originais pelos outputs foi crucial para despertar dúvida razoável na tech company sobre a deslinde do processo, capaz de ensejar proposta de acordo financeiramente vantajosa aos autores.

O relatório da USCO de 2025 corrobora essa abordagem, notando que datasets como o Common Crawl, com conteúdo pirata, podem diluir mercados se os outputs competirem no mercado, mas não define um padrão rígido para análise e aplicabilidade pelos tribunais sobre critérios a serem observados sobre diluição de mercado. Essa sensível divergência judicial evidente reflete a falta de consenso sobre como tratar a pirataria embutida no acesso as obras, um ponto crítico quando o scraping ignora fontes licenciadas versus piratas.

Igualmente a relevância da quantidade de conteúdo utilizado para fins de treinamento do modelo de IAG também parece distante de ser pacificado, uma vez que a quantidade (fator 3) tida como aceitável para caracterização do fair use também varia.

No caso Thomson Reuters a escala massiva foi vista como não razoável para o propósito alegado, enquanto no caso Anthropic foi justificada pela necessidade técnica de volume de dados para treinar LLMs. Essa discrepância sugere que tribunais podem estar inclinados a tolerar scraping extensivo se o objetivo for considerando transformador e volume necessário de dados para IAG, mas exigem evidências de controles, como filtros anti-memorização, citados na decisão proferida pelo juiz Chhabria no caso Bartz v. Anthropic.

Nesta esteia, a OCDE recomenda códigos de conduta para mitigar riscos éticos, mas os tribunais americanos ainda carecem de diretrizes uniformes, afinal, os aspectos éticos e legais do scraping, especialmente de fontes piratas, são subjacentes a todas as decisões. No caso *Bartz v. Anthropic* o juiz minimiza a pirataria como problema de curadoria, sendo admitido o scraping se o uso final não regurgitasse conteúdo original de obras utilizadas no treinamento, e no caso *Thomson Reuters* a violação por uso de contas de terceiros para scraping do conteúdo autoral reforçou a rejeição do fair use.

Isso contrasta com a posição da USCO que alerta para o acesso ilegal do conteúdo como fator contra a aplicação da doutrina do fair use, mas como seres sociais que somos enquanto seres humanos, os juízes naturalmente analisam os fatores a luz de suas convicções, mas mais do que isso observam a cultura, política e economia em que está sendo aplicada, o que torna latente o posicionamento “liberal” dos tribunais que possuem interesses político-econômicos ligados a IAG em virtude de seus tech hubs.

Além das decisões já analisadas, que delineiam os contornos iniciais da doutrina do fair use em relação ao data scraping e datasets mistos, persistem no sistema judiciário americano uma série de casos emblemáticos aguardando julgamento substantivo, os quais prometem aprofundar o debate sobre os direitos dos autores versus a inovação em IAG, tais como caso *The New York Times Co. v. Microsoft Corp. and OpenAI, Inc.* (Tribunal Distrital do Sul de Nova York, No. 1:23-cv-11195, ajuizado em dezembro de 2023, com foco no uso não autorizado de milhões de artigos jornalísticos para treinar o ChatGPT, inclusive burlando o paywall existente no site, questionando se tal scraping massivo constitui uso transformador ou dilui o mercado de conteúdo licenciado).

Outros casos pendentes incluem caso *Ziff Davis v. OpenAI* (ajuizado em abril de 2025 no Tribunal Distrital do Sul de Nova York), alegando infração por treinamento com artigos de marcas como IGN e Mashable, ou o caso *Music Publishers v. Anthropic PBC* (Tribunal Distrital do Sul de Nova York, agosto de 2025), envolvendo download ilegal de partituras piratas para o modelo Claude, com ênfase em violações direitos autorais digitais e fair use para outputs musicais, ambos aguardando julgamentos que poderão esclarecer a viabilidade de exceções para usos comerciais em datasets contaminados por pirataria.

Esses litígios, em estágios de conhecimento, representam não apenas uma extensão das tensões observadas nos precedentes anteriores, mas também uma oportunidade para os tribunais americanos delinearem normas mais precisas sobre transparência em curadoria de dados e

compensação aos autores, potencialmente influenciando reformas legislativas como o AI Training Data Transparency Act⁶ proposto no Congresso.

A divergência dos precedentes reflete variações jurisdicionais significativas, com tribunais de Delaware adotando uma postura protecionista aos direitos autorais, alinhada a indústrias tradicionais, rejeitando fair use em contextos comerciais diretos, enquanto os tribunais da Califórnia, mostram maior abertura à inovação, influenciados pela presença de empresas de como Meta e Anthropic, favorecendo transformatividade em pesquisa.

Essa divisão jurisdicional pode levar a um conflito nos tribunais de apelação (denominados “Circuit Courts”, análogos a 2ª instância brasileira), com apelações pendentes potencialmente escalando ao Supremo Tribunal americano, que poderia harmonizar o cenário nos Estados Unidos nos próximos anos, embora atualmente a falta de uniformidade persiste, desafiando autores e desenvolvedores a navegar em um cenário jurídico fragmentado.

PROBLEMÁTICA SISTEMICA DO MÉTODO PROBABILISTICO DA IAG

A IAG enquanto sistema estatístico de aprendizado, baseia-se em métodos probabilísticos para interpretar e reproduzir padrões linguísticos, visuais ou sonoros extraídos de grandes conjuntos de dados, cujas característica técnica, ao contrário da percepção intuitiva de que a “memoriza” ou “replica” diretamente as obras que consome, fundamenta-se em transformações matemáticas complexas, como a tokenização e o mapeamento vetorial de textos, imagens e sons.

Assim, no treinamento de modelos de linguagem, o processo envolve a conversão das palavras em unidades numéricas - tokens - que representam probabilidades de ocorrência de termos em determinados contextos, de forma que o modelo não armazena o texto integral, mas calcula a probabilidade de que uma sequência de palavras ocorra, a partir de um contexto estatístico aprendido.

Esse funcionamento descrito por Raffel (2023) no artigo denominado Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, explica que a IAG não realiza uma cópia literal das obras, mas extrai padrões linguísticos e semânticos que, combinados, permitem gerar novas respostas, formalmente inéditas, porém derivadas de uma corpora de dados

⁶ A legislação aprovada na Califórnia, *AI Training Data Transparency Act* (AB 1033, Califórnia, 2024), visa exigir que desenvolvedores de modelos de IAG divulguem publicamente, em seus sites, um resumo detalhado dos *datasets* usados no treinamento, incluindo fontes, *status* de licenciamento e inclusão de dados pessoais, para promover transparência e *accountability*, com vigência a partir de 1º de janeiro de 2026.

que pode conter obras protegidas.

Tal método probabilístico, embora tecnicamente não configure reprodução integral de conteúdo, levanta controvérsias quanto à violação do direito autoral, pois a abstração matemática não elimina o fato de que o aprendizado ocorre sobre obras humanas preexistentes.

A tokenização é, em essência, uma operação de leitura e fragmentação das obras originais, que permite ao sistema apreender o estilo, o vocabulário e a estrutura narrativa de cada texto, assim, ainda que o modelo não mantenha cópias acessíveis ou reconstruíveis das obras utilizadas, o processo de treinamento pode envolver cópias temporárias, cuja realização, sob a ótica da LDA, requer autorização prévia do titular.

Conforme apontado pelo relatório apresentado pelo Observatório de Direito Autoral (2025), o método matemático da IAG implica a existência de uma “memória probabilística” que reconfigura o conteúdo autoral em matrizes de peso sináptico, não sendo possível quantificar a contribuição específica de cada obra para o aprendizado do modelo, de forma com que essa indeterminabilidade técnica impede a mensuração direta de royalties e dificulta a aplicação de sistemas de compensação autoral baseados em uso individualizado.

A problemática adquire maior complexidade quando observada sob o prisma do direito brasileiro, cujo modelo autoralista personalista exige, como vimos, autorização expressa para qualquer forma de reprodução, ainda que parcial ou indireta, assim associada a problemática envolvendo a impossibilidade técnica de identificar a contribuição de cada obra no treinamento, inviabilizaria a aplicação dos mecanismos tradicionais de licenciamento e de fiscalização.

É patente que a abordagem probabilística da IAG desafia a própria noção jurídica de “reprodução”, já que os modelos operam sobre representações numéricas agregadas de tais obras, cuja arquitetura matricial, segundo a pesquisa do Observatório de Direito Autoral (2025), caracteriza-se pela compressão e distribuição da informação em parâmetros de rede neural, de modo que cada obra influencia o modelo apenas como vetor estatístico, e não como fonte de conhecimento de mundo como ocorre no pensamento humano orgânico.

Contudo, essa transformação das obras protegidas e vetores não neutraliza os efeitos econômicos da ausência de remuneração aos autores, cujos textos, músicas ou imagens que integram datasets podem vir a sofrer diluição de mercado em determinadas circunstâncias, uma vez que o modelo reproduz estilos, formatos e estruturas expressivas derivadas de suas criações, podendo levar a substituição funcional da obra original no consumo cultural, o que atribui complexidade adicional a IAG que embora opere sem “memória literal” por sua natureza

matemática, ainda pode ser entendida como ferramenta apta a gerar externalidades negativas no mercado, competindo com obras protegidas.

Os impactos das violações de direitos autorais no treinamento de IAG podem se manifestar em múltiplas dimensões, com repercussões econômicas, jurídicas e sociais que transcendem o mero uso técnico de obras protegidas.

Economicamente, a incorporação não autorizada de obras em datasets poderia ser apta resulta em diluição de mercado para os autores originais, onde modelos de IAG capazes de replicar estilos, estruturas ou conteúdos semelhantes competem diretamente com as criações humanas, reduzindo a demanda por licenças e royalties. Conforme o relatório da USCO (2025), essa substituição funcional pode levar a perdas financeiras estimadas em bilhões de dólares anuais para indústrias criativas, como jornalismo e literatura, como o caso *The New York Times v. OpenAI* ilustrando uma queda de aproximadamente 20% na audiência de conteúdos pagos, supostamente devido a outputs gerados por IA, cujo efeito similar é alegado no Brasil no caso *Folha de São Pulo v. OpenAI*. Essa dinâmica econômica desbalanceada agrava a concentração de poder nas grandes empresas de tecnologia, que monetizam modelos comerciais sem compensação proporcional aos autores dos insumos, perpetuando desigualdades no ecossistema cultural.

Juridicamente, as violações geram insegurança regulatória e um aumento exponencial de litígios, como evidenciado pelo mapeamento de mais de 50 processos (apurados até setembro de 2025) somente nos Estados Unidos, muitos envolvendo datasets possivelmente contaminados por pirataria, denominadas *shadow libraries*. No Brasil, a rigidez da LDA e a ausência de exceções flexíveis como o *fair use* intensificam esses impactos, possibilitando entendimentos equivocados de que práticas como *web scraping* poderiam ser entendidas como contrafação potencial (art. 184 do CP), com riscos de sanções civis e penais que desestimulariam investimentos locais em IAG.

Socialmente, essas violações ameaçam a sustentabilidade da produção cultural humana, desincentivando criadores ao deteriorar significativamente sua remuneração e autonomia, o que pode resultar em uma massificação de conteúdos gerados por IAG baseados em padrões estatísticos derivados de obras não autorizadas.

No contexto global, como destacado pela OMPI, isso acentua assimetrias entre países desenvolvidos (com forte indústria de IA) e emergentes como o Brasil, onde a dependência de tecnologias estrangeiras agrava a vulnerabilidade cultural, sendo notório que os impactos não se limitam a prejuízos individuais, mas configuram uma ameaça sistêmica ao equilíbrio entre inovação tecnológica e preservação da criatividade humana, exigindo soluções híbridas como

licenciamento coletivo e outros institutos para mitigar riscos sem frear o progresso.

No âmbito regulatório, o PL nº 2.338/2023 avança timidamente ao reconhecer a necessidade de transparência e compensação no uso de obras para treinamento de IA, mas permanece aquém das exigências técnicas do método probabilístico, já que o texto legal pressupõe que os desenvolvedores são capazes de identificar e documentar as fontes utilizadas, exigência inviável em modelos treinados sobre datasets massivos e parcialmente abertos, nos quais a distinção entre conteúdo licenciado, de domínio público e protegido é estatisticamente diluída.

Ressalta-se que o método probabilístico não permite mensurar o grau de contribuição de uma obra específica, tornando impossível calcular uma remuneração justa sem adotar fórmulas baseadas em métricas estimadas, como frequência ou relevância semântica.

A discussão sobre a responsabilização autoral no contexto da IAG, portanto, exige uma reformulação conceitual, uma vez que o modelo tradicional de direito autoral, centrado na obra individualizável e no autor determinado, mostra-se insuficiente diante da natureza estatística e difusa do aprendizado de máquina.

Sendo assim, a problemática sistêmica do uso de obras protegidas no contexto do método probabilístico da IAG revela a incompatibilidade estrutural entre a natureza coletiva, estatística e automatizada do aprendizado de máquina e o modelo jurídico personalista do direito autoral brasileiro, e a superação dessa tensão requer soluções normativas e técnicas conjuntas, que reconheçam a singularidade do processo de aprendizado algorítmico sem abdicar da proteção dos direitos dos criadores humanos, de forma que a regulação deve avançar para além da lógica da autorização individual e incorporar múltiplos mecanismos que se aplicados conjuntamente poderão ser capazes de transformar a tensão entre inovação e proteção autoral em um equilíbrio dinâmico e sustentável.

PERSPECTIVAS GLOBAIS E CAMINHOS POSSÍVEIS DE HARMONIZAÇÃO

A fragmentação regulatória em torno da IAG e dos direitos autorais constitui certamente um dos maiores desafios contemporâneos da governança global da propriedade intelectual, embora tratados multilaterais clássicos, como a Convenção de Berna e o TRIPS, fixem padrões mínimos de proteção autoral obrigatórios para os Estados signatários, ambos foram concebidos em contexto tecnológico anterior à automação criativa e, portanto, não contemplam o uso de obras

protegidas em datasets de treinamento de sistemas de IAG.

A ausência de normas específicas uniformes obriga cada país a interpretar isoladamente a extensão de suas normas autorais diante dos processos de aprendizado de máquina, o que gera sobreposições, lacunas e conflitos normativos. Essa dificuldade se amplifica na realidade hiperconectada verificada na IAG, em que os elementos relevantes do processo tecnológico podem estar dispersos em distintas jurisdições: a empresa exploradora pode ter sede em um país e operar globalmente pela web; os data centers que armazenam e processam dados localizam-se em diferentes continentes; os datasets são formados por conteúdos hospedados em sites distribuídos pelo mundo; e os engenheiros ou prestadores de serviço que colaboram no treinamento podem ser cidadãos de outras nações. Essa multiplicidade de pontos de conexão - geográfica, técnica e jurídica - evidenciam o desafio de estabelecer um critério único de responsabilidade e de determinar a legislação aplicável aos atos praticados no processo de aprendizado de sistemas de IA.

A Organização Mundial da Propriedade Intelectual (OMPI) tem se posicionado como principal foro de debate sobre a matéria. Desde 2019, por meio da iniciativa “WIPO Conversation on Intellectual Property and Artificial Intelligence”, a entidade conduz consultas públicas e encontros técnicos sobre temas como autoria e titularidade de obras geradas por IA, proteção de bases de dados e requisitos de transparência de datasets. A entidade reconhece a insuficiência dos instrumentos internacionais vigentes e propõe o estudo de mecanismos de licenciamento coletivo e de exceções limitadas para mineração de dados. Contudo, até o momento não houve consenso entre os Estados-membros, refletindo divergências de natureza tanto econômica quanto jurídica.

Países com forte indústria tecnológica, como Estados Unidos, Japão e Coreia do Sul, tendem a privilegiar modelos regulatórios que favoreçam inovação e liberdade de uso de dados, enquanto blocos como a União Europeia e o Reino Unido têm enfatizado a proteção dos titulares de direitos e a exigência de transparência no treinamento de modelos generativos.

Nos países em desenvolvimento, incluindo o Brasil, Índia e África do Sul, a questão assume contornos distintos, uma vez que essas economias buscam equilibrar a proteção de seus setores culturais com a necessidade de acesso a tecnologias e dados para o desenvolvimento de capacidades nacionais em IAG.

Relatórios das reuniões do Comitê de Desenvolvimento e Propriedade Intelectual (CDIP) da OMPI (2023) e da UNESCO (2024) indicam que regulações excessivamente restritivas podem acentuar a dependência tecnológica de países periféricos, ao passo que regimes demasiadamente

permissivos tendem a fragilizar a sustentabilidade econômica de seus criadores. Este dilema é particularmente sensível para o Brasil, que possui sistema autoral personalista e ainda carece de infraestrutura nacional de dados e políticas de licenciamento coletivo voltadas para o uso de obras em IAG.

Diante da dificuldade de alcançar consenso multilateral, observa-se crescente movimento em direção a soluções contratuais e setoriais, como é possível identificar nos acordos privados entre desenvolvedores de IAG e detentores de grandes acervos de conteúdo, como os acordos firmados pela OpenAI com a Financial Times, Shutterstock, Reddit, Axel Springer, Atlantic e Vox Media, bem como iniciativas de licenciamento do Google com Reddit ou parceria da Runway com Getty Images para desenvolvimento de modelo de geração de vídeo para ser aperfeiçoado (fine tuning) para uso comercial.

Embora tais acordos certamente envolvam compensação financeira e previsibilidade jurídica, sua natureza confidencial e a concentração de benefícios em grandes grupos empresariais suscitam questionamentos sobre equidade e representatividade.

Outra tendência emergente é o desenvolvimento de licenças abertas específicas para IA, tal como as Open Responsible AI Licenses criadas em 2023 e os projetos AI Data License da Creative Commons que propõem estruturas contratuais que permitem aos autores definir, de modo granular, se e em que condições suas obras podem ser utilizadas para treinamento de modelos. Essas soluções, embora relevantes para aumentar a clareza jurídica e reduzir litígios, ainda são incipientes e carecem de padronização internacional e, neste contexto, modelos de licenciamento coletivo também ganham espaço como proposta de harmonização intermediária, inspirados em sistemas já consolidados para música e audiovisual, tais modelos poderiam ser geridos por entidades de gestão coletiva que negociariam licenças amplas com desenvolvedores de IAG e redistribuiriam os valores arrecadados de forma proporcional aos titulares.

Inclusive, em audiência pública realizada na Câmara dos Deputados no contexto da Comissão Especial sobre Inteligência Artificial diversos convidados defendem que essa solução poderia reduzir custos de transação e garantir remuneração aos autores sem comprometer a viabilidade técnica do treinamento de modelos de larga escala, trazendo o exemplo entidade americana Corporate Clearance Center (CCC) que possui a AI Systems Training License, em funcionamento e com adesão voluntária.

A busca por um regime internacional de harmonização ainda parecer estar longe da realidade, pois passa pela necessidade de alinhar princípios de transparência, compensação a

indústria criativa (de algum modo) e uso ético de dados com as realidades técnicas do aprendizado de máquina.

A experiência comparada demonstra que nenhuma jurisdição teve êxito, até o momento, em conciliar plenamente esses elementos complexos decorrentes da nova tecnologia e relações de política econômica dos países, afinal é notório a aposta mundial na tecnologia advinda da como nova externalização de poder e influência da nação sob outros territórios.

A harmonização global futura efetiva, embora não esteja em curto horizonte, far-se-á necessária, exigirá instrumentos internacionais complementares, capazes de atualizar os tratados de Berna e TRIPS à era algorítmica, incorporando cláusulas sobre mineração de dados, deveres de informação e mecanismos de compensação, mas até que tal consenso se consolide, o cenário global permanecerá fragmentado, marcado pela convivência de regimes autorais nacionais e acordos privados que refletem, antes de tudo, o poder econômico e tecnológico de seus signatários.

CONSIDERAÇÕES FINAIS

As transformações promovidas pela IAG colocam o direito autoral diante de um de seus maiores desafios históricos, pois a arquitetura probabilística que sustenta o aprendizado desta tecnologia rompe com a lógica clássica de autoria individual e de reprodução material da obra, fundamentos sobre os quais o sistema autoral brasileiro foi construído.

A pesquisa demonstrou que há uma incompatibilidade estrutural entre o modelo personalista da LDA (Lei nº 9.610/1998) e o processo de treinamento das IAGs, que opera de forma coletiva, difusa e automatizada. Essa incompatibilidade não implica a impossibilidade de tutela jurídica, mas exige uma reinterpretação dos conceitos tradicionais de uso, reprodução e originalidade à luz da realidade tecnológica contemporânea. A simples aplicação das categorias jurídicas existentes, sem adequação aos processos de aprendizado probabilístico, tende a produzir insegurança normativa e a dificultar tanto a proteção dos autores quanto o desenvolvimento tecnológico nacional.

Nesse cenário, a principal contribuição deste artigo é evidenciar que a proteção das obras humanas e o avanço da inteligência artificial não são objetivos mutuamente excludentes. Ao contrário, a harmonização entre ambos depende da adoção de instrumentos jurídicos e técnicos complementares, capazes de equilibrar os direitos individuais dos criadores com o interesse

coletivo na inovação. Afinal, a proteção ao direito autoral constitucionalmente previsto não é absoluto, devendo ser equilibrado com princípios constitucionais da função social da propriedade, do acesso a cultura, a informação e a educação (arts. 5, XIV; 6; 215 e 218 da CF).

A pesquisa aponta que medidas de transparência, licenciamento coletivo e salvaguardas técnicas contra memorização literal podem coexistir e formar a base de um novo paradigma regulatório, mais compatível com a natureza estatística e globalizada do aprendizado de máquina. Esse equilíbrio exige abandonar soluções binárias - proibição ou liberalização absoluta - e construir uma abordagem intermediária, de cooperação entre Estado, empresas e titulares de direitos.

No Brasil, em razão da inexistência de uma cláusula aberta de uso legítimo ou transformativo, a tendência é que o treinamento técnico de modelos com obras protegidas, mesmo por meio de tokenização, exija autorização expressa do titular, salvo demonstração de compatibilidade com os parâmetros da Regra dos Três Passos. A ausência dessa análise normativa específica cria incerteza regulatória e coloca o país em desvantagem comparativa frente a sistemas jurídicos que já preveem exceções voltadas à pesquisa e inovação.

No plano normativo em discussão no país, a implementação dos artigos sobre direito autoral em projetos de lei que tratam sobre IAG podem constituir o primeiro passo concreto rumo ao futuro da IA no Brasil, requerendo uma escolha política social e econômica norteadora, e mais do que isso, um olhar atento ao mundo real em que a legislação estará inserida, em que não será possível fazer contato autor por autor, editora por editora, seja por temas de escala, como por aspectos de custo de transação elevados.

A regulamentação também deve prever métricas estimativas para remuneração proporcional dos autores, se e quando cabível, diante da inviabilidade técnica de rastrear a contribuição individual de cada obra em modelos treinados com bilhões de parâmetros. Indicadores como frequência de uso, similaridade semântica ou peso estatístico das fontes podem servir como proxies razoáveis para calcular compensações. O objetivo não é obter precisão matemática, mas garantir um critério previsível e verificável que promova justiça distributiva. Assim, o licenciamento coletivo digital surge como alternativa viável, reduzindo custos de transação e assegurando que autores de menor visibilidade também sejam compensados pelo uso de suas obras, apesar da necessidade de ponderação sobre a posição do Brasil ao fazer sua escolha legislativa.

No campo técnico, é fundamental que as políticas de governança de IA incluam

compromissos de mitigação de memorização literal e de reprodução não intencional de obras protegidas, tais como filtros de verificação de similaridade, citação obrigatória de fontes, logs de treinamento e auditorias independentes de datasets devem integrar práticas obrigatórias de compliance de IA, de forma que tais ferramentas não apenas previnem infrações autorais, mas também aumentam a confiabilidade e a segurança dos sistemas de IAG.

A análise comparada realizada neste trabalho demonstra, ainda que preliminarmente, que não há consenso internacional sobre o enquadramento jurídico do treinamento de IAG. Os Estados Unidos adotam um modelo casuístico baseado no fair use, cuja cláusula aberta estimula fortemente a indústria da IAG e desenvolvimento tecnológico, sem contudo deteriorar o mercado original das obras.

Diante desse panorama, o Brasil tem a oportunidade de desenvolver um modelo híbrido, que combine exceções limitadas para pesquisa e inovação com um regime de licenciamento coletivo e transparência obrigatória para usos comerciais, sendo certo que essa solução permitiria ao país alinhar-se às boas práticas internacionais, preservando sua tradição personalista de proteção ao autor. A escolha regulatória do Brasil frente a problemática do treinamento da IAG será fundamentalmente importante para o progresso tecnológico e econômico do país nas próximas décadas, especialmente dado o cenário econômico ambiental brasileiro que nos colocaria como país de matriz energética limpa para fins de instalação de data centers, o que poderia vir a ser desincentivado por uma legislação excessivamente protecionista.

O caso *Folha de S. Paulo v. OpenAI*, analisado no presente artigo, representa o ponto de inflexão dessa transição, em que independentemente do desfecho judicial, o processo inaugura o debate sobre a aplicação prática do direito autoral brasileiro às tecnologias de inteligência artificial e deverá servir de parâmetro para futuras regulamentações e contratos privados. Em contexto tão fluido e dinâmico da tecnologia, o acompanhamento sistemático de litígios semelhantes, tanto no Brasil quanto no exterior, torna-se essencial para observar padrões jurisprudenciais, compreender os efeitos econômicos das decisões e orientar o aperfeiçoamento normativo.

As perspectivas futuras indicam a necessidade de ampliar a cooperação entre o direito e as ciências tecnológicas, pesquisas multidisciplinares que integrem juristas, engenheiros de dados e especialistas em ética da IA são indispensáveis para construir soluções regulatórias efetivas e tecnologicamente viáveis.

O diálogo entre direito e tecnologia permitirá desenvolver métricas objetivas de

rastreabilidade, ferramentas de auditoria e modelos preditivos de impacto autoral, promovendo um ambiente de inovação responsável. Além disso, a criação de observatórios permanentes de litígios e de políticas públicas em IA pode auxiliar na coleta sistemática de dados, no monitoramento das decisões judiciais e na formulação de recomendações legislativas baseadas em evidências.

Por fim, este artigo reafirma que a IAG, longe de representar uma ameaça ao direito autoral, deve ser compreendida como oportunidade para sua evolução, a incorporação de princípios de transparência e compartilhamento de conhecimento podem fortalecer a função social da propriedade intelectual, transformando o direito autoral em instrumento de equilíbrio entre criatividade humana e inovação tecnológica, propiciando disseminação de cultura e pluralidade de ideias.

A convergência entre ética, ciência e direito poderá consolidar um novo paradigma, no qual o progresso tecnológico e a valorização da criação intelectual se complementem, promovendo um ecossistema cultural mais inclusivo, sustentável e comprometido com o desenvolvimento humano.

REFERÊNCIAS

ALMEIDA, Silvia F. de; TAKAHASHI, Gabriel S.; MOREIRA, Miguel S.; OLIVEIRA, Lara de. **Impactos do PL 2338/2023: mensuração da redução de ganhos de produtividade decorrente da remuneração de direitos autorais proposta sobre o uso de IA**. Observatório de Direitos Autorais e Tecnologia, 2025. Disponível em: <https://observatorio.cc/wp-content/uploads/2025/09/Ecoa-Remuneracao-e-IA_20250908-1.pdf> Acesso em: 4 nov. 2025.

ANTHROPIC. **Constitutional AI: Harmlessness from AI feedback**. San Francisco: Anthropic Research, 2023. Disponível em: <<https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback>> Acesso em: 4 nov. 2025.

O'BRIEN, Matt. **OpenAI and the Associated Press announce partnership on generative AI**. *Associated Press (AP)*, Nova York, 29 jul. 2023. Disponível em: <<https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a>> Acesso em: 4 nov. 2025.

BARBOSA, Denis Borges. **Direito de Autor: Questões fundamentais de direito de autor**. Ebook. 3. ed. Rio de Janeiro: Lumen Juris, 2012.

BARBOSA, Denis Borges. **Introdução à propriedade intelectual**. Ebook. 2. ed. Rio de Janeiro: Lumen Juris; DBBA, 2010. Disponível em: <https://www.dbba.com.br/wp-content/uploads/introducao_pi.pdf>. Acesso em: 4 nov. 2025.

BASSO, Maristela. **A propriedade intelectual e a proteção do software**. Revista CEJ, Brasília, n. 21, p. 16–30, abr./jun. 2003. Disponível em: <<https://revistacej.cjf.jus.br/cej/index.php/revcej/article/view/541/721>>. Acesso em: 4 nov. 2025.

BASSO, Maristela. **As exceções e limitações aos direitos do autor e a observância da regra do teste dos três passos (three-step test)**. Revista da Faculdade de Direito, Universidade de São Paulo, São Paulo, 2007. Disponível em: <https://www.researchgate.net/publication/269817259_As_excecoes_e_limitacoes_aos_direitos_do_autor_e_a_observancia_da_regra_do_teste_dos_tres_passos_three-step-test> Acesso em: 4 nov. 2025.

BSBC ADVOGADOS. **A evolução dos tratados internacionais sobre direito autoral: reflexos do antagonismo entre droit d’auteur e copyright**. São Paulo: BSBC Advogados, 2025. Disponível em: <<https://bsbcadvogados.com.br/a-evolucao-dos-tratados-internacionais-sobre-direito-autoral-reflexos-do-antagonismo-entre-droit-dauteur-e-copyright/>>. Acesso em: 4 nov. 2025.

BRASIL. **Lei nº 9.610, de 19 de fevereiro de 1998**. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. Diário Oficial da União, Brasília, DF, 20 fev. 1998. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l9610.htm> Acesso em: 4 nov. 2025.

BRASIL. **Câmara dos Deputados. Projeto de Lei nº 2.338, de 2023**. Dispõe sobre o uso da Inteligência Artificial. Brasília, DF: Câmara dos Deputados, 2023. Disponível em: <<https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2338898>>. Acesso em: 4 nov. 2025.

BRASIL. **Senado Federal. Projeto de Lei nº 2.338, de 2023**. Dispõe sobre o uso da Inteligência Artificial. Autor: Senador Rodrigo Pacheco (PSD/MG). Brasília, DF: Senado Federal, 2023. Disponível em: <<https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>>. Acesso em: 4 nov. 2025.

BRASIL. **Tribunal de Justiça. Processo nº 1107237-96.2025.8.26.0100**. Empresa Folha da Manhã S.A. (**requerente**) v. OpenAI Global LLC; OpenAI GP LLC; OpenAI Inc. (requeridos). Procedimento Comum Cível – Concorrência desleal. São Paulo: 3ª Vara Empresarial e de Conflitos de Arbitragem, Foro Central Cível, 2025. Disponível em: <https://internetlab.org.br/wp-content/uploads/2025/08/doc_141965902.pdf> Acesso em 4 nov. 2025.

CASTRO, Lincoln Antonio de. **Noções sobre direito autoral**. Rio de Janeiro: Ministério Público do Estado do Rio de Janeiro (MPRJ), 2018. Disponível em: <https://www.mprj.mp.br/documents/20184/2813639/Lincoln_Antonio_de_Castro.pdf>. Acesso em: 4 nov. 2025.

COMMON CRAWL. **The data – FAQ**. San Francisco: Common Crawl, 2025. Disponível em: <<https://commoncrawl.org/faq>>. Acesso em: 4 nov. 2025.

CORREIA, Fernanda Nunes. **A inteligência artificial generativa e o direito autoral: desafios à luz da Lei nº 9.610/1998**. Revista Foco, v. 18, n. 3, 2025. Disponível em: <<https://ojs.focopublicacoes.com.br/foco/article/view/9143>>. Acesso em: 4 nov. 2025.

DEUTSCHE UNESCO KOMMISSION. **AI in the cultural and creative industries**. Bonn: Deutsche UNESCO-Kommission, nov. 2024. Disponível em: <https://www.unesco.de/assets/dokumente/Deutsche_UNESCO-Kommission/02_Publikationen/DUK_Broschuere_KI_und_Kultur_EN_web_02.pdf>. Acesso em: 4 nov. 2025.

EBSCO RESEARCH STARTERS. **Copyright Term Extension Act**. Ipswich: EBSCO, 2024. Disponível em: <<https://www.ebsco.com/research-starters/political-science/copyright-term-extension-act>>. Acesso em: 4 nov. 2025.

EUROPEAN UNION. **Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market**. Official Journal of the European Union, 2019. Disponível em: <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L0790>>. Acesso em: 6 out. 2025.

FRAZÃO, Ana. **Direitos autorais e inteligência artificial**. Blog GEN Jurídico, São Paulo, 12 abr. 2024. Disponível em: <<https://blog.grupogen.com.br/juridico/areas-de-interesse/direito-digital/direitos-autorais-inteligencia-artificial/>>. Acesso em: 4 nov. 2025.

IBM – INTERNATIONAL BUSINESS MACHINES CORPORATION. **Dataset – conceitos e aplicações**. IBM Think Brasil, 2025. Disponível em: <<https://www.ibm.com/br-pt/think/topics/dataset>>. Acesso em: 4 nov. 2025.

INTERPOL. **Digital piracy. Crimes: Illicit Goods – Shop Safely**. Lyon: INTERPOL, 2025. Disponível em: <<https://www.interpol.int/Crimes/Illicit-goods/Shop-safely/Digital-piracy#:~:text=Pirataria%20digital%20refere%2Dse%20%C3%A0,por%20direitos%20autorais%20pela%20internet>>. Acesso em: 4 nov. 2025.

KAPLAN, Daniel. **Artificial intelligence, copyright, and authorship: comparative perspectives on originality**. SSRN Electronic Journal, 2025. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5253011&download=yes>. Acesso em: 4 nov. 2025.

KAUFMAN, Dora. **Desmistificando a inteligência artificial**. Belo Horizonte: Autêntica, 2022.

LEMOS, Ronaldo. **Direito, tecnologia e cultura**. Ebook. Rio de Janeiro: FGV Editora, 2005.

LEMOS, Ronaldo. **Inteligência artificial, cultura e direitos autorais**. Ebook. Revista FGV Direito Rio, Rio de Janeiro, 2024.

MADIGAN, Kevin. **Generative AI copyright infringement: lessons from past fair use cases**. Washington, D.C.: Hudson Institute, 2024. Disponível em: <<https://www.hudson.org/information-technology/generative-ai-copyright-infringement-lessons-past-fair-use-cases>>. Acesso em: 4 nov. 2025.

MARANHÃO, Juliano. **Apresentação na Audiência Pública da Comissão Especial sobre Inteligência Artificial (PL 2.338/2023)**, Câmara dos Deputados, 25 set. 2025. Brasília: Câmara dos Deputados, 2025. Disponível em: <<https://www2.camara.leg.br/atividade-legislativa/comissoes/comissoes-temporarias/especiais/57a-legislatura/comissao-especial-sobre-inteligencia-artificial-pl-2338-23/apresentacoes-em-eventos/JULIANOMARANHOProfessordaUSPAudienciaCamaraDeputadosSet25.pdf>>. Acesso em: 4 nov. 2025.

MITCHELL, Margaret; WU, Simone; ZALDIVAR, Andrew; BARNES, Parker; VASSERMAN, Lucy; HUTCHINSON, Ben; SPITZER, Elena; RAJI, Inioluwa Deborah; GEBRU, Timnit. **Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency**, Atlanta, EUA, 29–31 jan. 2019. New York: ACM, 2019. Disponível em: <<https://arxiv.org/pdf/1810.03993>>. Acesso em: 4 nov. 2025.

MILLER, Olivia. **The importance of data curation: a step-by-step guide**. Medium, 2025. Disponível em: <<https://medium.com/@oliviamiller048/the-importance-of-data-curation-a-step-by-step-guide-d622ef23568>>. Acesso em: 4 nov. 2025.

MITCHELL, Ryan. **Web scraping with Python**. Ebook. 2. ed. Sebastopol: O'Reilly Media, 2018.

NVIDIA. **Tokens de IA explicados**. Blog NVIDIA Brasil, 2025. Disponível em: <<https://blog.nvidia.com.br/blog/tokens-ia-explicados/>>. Acesso em: 4 nov. 2025.

OCDE. **Intellectual property issues in artificial intelligence trained on scraped data**. Paris: OECD Publishing, 2025. fl. 19. Disponível em: <https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/intellectual-property-issues-in-artificial-intelligence-trained-on-scraped-data_a07f010b/d5241a23-en.pdf>. Acesso em: 4 nov. 2025.

RAFFEL, Colin; SHAZEER, Noam; ROBERTS, Adam; LEE, Katherine; NARANG, Sharan; MATENA, Michael; ZHOU, Yanqi; LI, Wei; LIU, Peter J. **Exploring the limits of transfer learning with a unified text-to-text transformer**. [S. l.]: arXiv, set. 2020. DOI: 10.48550/arXiv.1910.10683. Disponível em: <<http://arxiv.org/abs/1910.10683>>. Acesso em: 4 nov. 2025.

RAMOS, P. H.; BARRETO, J.; GARROTE, M. **Remuneração por Direitos Autorais em IA: Limites e Desafios de Implementação**. Policy Briefs Reglab, n. 3. São Paulo: Reglab, 2025. Disponível em: <<https://reglab.com.br/remuneracao-por-direitos-autorais-em-ia-limites-e-desafios-de-implementacao/>>. Acesso em: 4 nov. 2025.

RIBEIRO, N.; GARROTE, M. **Mapeamento global de leis de direitos autorais e exceções para mineração de dados**. Contexto Reglab, n. 1. São Paulo: Reglab, 2025. Disponível em: <<https://reglab.com.br/mapeamento-global-de-leis-de-direitos-autorais-e-excecoes-para-mineracao-de-dados/>>. Acesso em: 4 nov. 2025.

RUSSELL, Stuart; NORVIG, Peter. **Artificial intelligence: a modern approach**. Ebook. 4. ed. Boston: Pearson, 2020.

SANTOS, Laura Alice Bogéa Praseres. **A tutela jurídica da inteligência artificial e os limites da responsabilidade civil: um estudo sobre a imputação de danos por decisões automatizadas**. 2023. Dissertação – Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP), Brasília, 2023. Disponível em: <https://repositorio.idp.edu.br/bitstream/123456789/3173/1/DISSERTA%C3%87%C3%83O_LAURA%20ALICE%20BOG%C3%89A%20PRASERES%20SANTOS_MESTRADO%20EM%20DIREITO.pdf>. Acesso em: 4 nov. 2025.

SOKLER, Bruce; HECHT, Alexander; FJELD, Christian Tamotsu; TIKHONOVSKY, Matthew. **US Copyright Office releases third report on AI copyrightability, and House Committee advances AI legislation**. AI: The Washington Report. ML Strategies Viewpoints, Washington, D.C., 16 maio 2025. Disponível em: <<https://www.mlstrategies.com/insights-center/viewpoints/54031/2025-05-16-us-copyright-office-releases-third-report-ai#:~:text=Market%20dilution%20occurs%20when%20%E2%80%9Ca,as%20in%20their%20training%20data.%E2%80%9D>>. Acesso em: 4 nov. 2025.

TRINDADE, Isadora. **Pirataria digital à luz do direito brasileiro**. Mackenzie Legal Hackers, São Paulo, 20 fev. 2024. Disponível em: <https://mackenzie.legalhackers.org/2024/02/pirataria-digital-a-luz-do-direito-brasileiro/>. Acesso em: 4 nov. 2025.

UNIÃO EUROPEIA. **Diretiva (UE) 2019/790 do Parlamento Europeu e do Conselho, de 17 de abril de 2019**. Relativa aos direitos de autor e direitos conexos no mercado único digital. Jornal Oficial da União Europeia, Bruxelas, 2019. Disponível em: <<https://eur-lex.europa.eu/eli/dir/2019/790/oj?locale=pt>>. Acesso em: 4 nov. 2025.

UNITED STATES. **The Constitution of the United States of America**. Philadelphia: National Constitution Center, 2024. Disponível em: <<https://constitutioncenter.org/media/files/Port-Constitution%208-19.pdf>>. Acesso em: 4 nov. 2025.

UNITED STATES COPYRIGHT OFFICE. **Chapter 300 – Copyrightable authorship**. Washington, D.C.: U.S. Copyright Office, 2024. Disponível em: <<https://www.copyright.gov/comp3/chap300/ch300-copyrightable-authorship.pdf>>. Acesso em: 4 nov. 2025.

UNITED STATES COPYRIGHT OFFICE. **Copyright and Artificial Intelligence – Part 2: Copyrightability Report**. Washington, D.C.: U.S. Copyright Office, 2024. Disponível em: <<https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf>>. Acesso em: 4 nov. 2025.

UNITED STATES COPYRIGHT OFFICE. **Copyright and Artificial Intelligence – Part 3: Generative AI Training Report**. Washington, D.C.: U.S. Copyright Office, 2025. Disponível em: <<https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>>. Acesso em: 4 nov. 2025.

WIPO – WORLD INTELLECTUAL PROPERTY ORGANIZATION. **Issues Paper on Intellectual Property Policy and Artificial Intelligence**. Geneva: WIPO, 2020. Disponível em: <<https://www.wipo.int/en/web/frontier-technologies/artificial-intelligence/conversation>>. Acesso em: 6 out. 2025.

WORLD INTELLECTUAL PROPERTY ORGANIZATION (WIPO). **Artificial intelligence. Geneva:** WIPO, 2024. Disponível em: <<https://www.wipo.int/en/web/frontier-technologies/artificial-intelligence/index#:~:text=Copyright%20infrastructure%2C%20the%20generally%20unseen,and%20across%20different%20creative%20industries>>. Acesso em: 4 nov. 2025.

ZUCOLOTO, Graziela Ferrero; NOGUEIRA, José Ricardo Fucidji Junior. **O uso da “Regra dos Três Passos” (Three Step Test) como mecanismo de flexibilização do direito autoral.** IntersciencePlace, v. 17, n. 4, p. 1-18, 2022. Disponível em: <<https://www.interscienceplace.org/index.php/isp/article/download/350/78/531>>. Acesso em: 12 out. 2025.