Remuneração por Direitos Autorais em IA

Limites e Desafios de Implementação

reglab
centro de estratégia
& regulação

Sobre o Reglab

O Reglab é um *think tank* especializado em pesquisa e consultoria que auxilia empresas, associações empresariais e formuladores de políticas no planejamento orientado por dados e análises de impacto. Nosso foco está na tomada de decisões responsáveis e estratégicas, desvendando os desafios regulatórios do setor de mídia e tecnologia.

Nosso objetivo é promover pesquisas baseadas em evidências que aumentem a responsabilidade e estabeleçam marcos e metas significativas para o ecossistema.

Saiba mais em www.reglab.com.br

Sobre a Série Policy Briefs

A Série *Policy Briefs* engloba estudos que avaliam políticas públicas existentes ou propostas, utilizando dados qualitativos e quantitativos para informar e orientar decisões estratégicas. O objetivo é trazer questões complexas de forma acessível, destacando os principais pontos de análise, impactos e possíveis recomendações.

Expediente

Diretor Executivo: Pedro Henrique Ramos

Coordenadora de Pesquisa: Marina Garrote

Autores(as): Pedro Henrique Ramos, Julia de Albuquerque Barreto, Marina Garrote

Pesquisadoras: Stephanie Mathias de Souza

Diagramação Final: Eliza Natsuko Shiroma

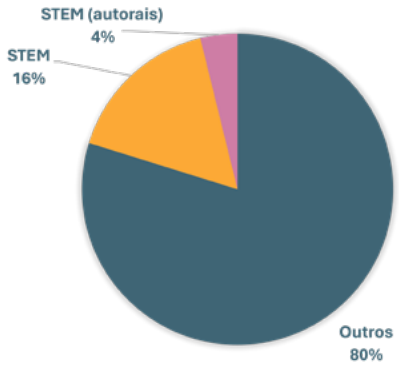
Citação Sugerida: RAMOS, P. H.; BARRETO, J.; GARROTE, M. *Remuneração por Direitos Autorais em IA: Limites e Desafios de Implementação.* Policy Briefs Reglab, n. 3. São Paulo: Reglab, 2025.

Sumário Executivo

O debate sobre direitos autorais e Inteligência Artificial Generativa (IAG) é um dos grandes assuntos do momento na regulação, com diferentes perspectivas e opiniões. O Reglab resolveu enfrentar esse tema a partir de uma **perspectiva pouco valorizada no debate regulatório: a visão de profissionais das áreas de STEM** (ciências exatas, tecnologia, engenharia e matemática).

Para isso, entrevistamos **cientistas da computação, engenheiros de software, especialistas em aprendizado de máquina e professores universitários**, com o objetivo de compreender de que forma o treinamento de modelos de IAG envolve o uso de conteúdo protegido por direitos autorais, e quais são os desafios técnicos para viabilizar propostas de remuneração associadas a esse uso.

Essa é uma pesquisa inédita no Brasil. Para se ter ideia da importância desses dados, examinamos as 24 audiências da CTIA - Comissão Temporária Interna sobre Inteligência Artificial no Brasil, que analisou o projeto de lei n. 2.338/23, e observamos a baixa participação desses profissionais no debate sobre IAG e direitos autorais.



Perfil de participantes e contribuições da CTIA - Comissão Temporária Interna sobre IA

Nuvem de palavras gerada por meio do software Wordclouds.com, a partir de informações geradas automaticamente pelo Atlas.ti, utilizando a ferramenta "Concepts".

Entre os principais achados da pesquisa, destacamos:

- a seleção de dados é um processo complexo e crítico para os modelos de IA, e **não só a quantidade, mas também a qualidade e diversidade dos dados** influenciam de maneira determinante no desempenho dos modelos;

- embora existam abordagens técnicas que permitem rastrear o fluxo e a origem dos dados usados no treinamento, os entrevistados indicaram que **ainda não há soluções escaláveis e confiáveis para medir a contribuição específica de cada obra em modelos de larga escala**. Isso não parece ser uma escolha de mercado, mas uma limitação estrutural da tecnologia atual, especialmente no aprendizado de máquina;

- isso acontece porque **modelos baseados em aprendizado de máquina não armazenam dados como um banco de referência consultável,** mas sim como padrões vetoriais, generalizados a partir de probabilidades estatísticas, “quebrando” dados e convertendo as informações obtidas em números. Assim, **determinar o impacto exato de cada obra no modelo final é, na prática, impossível.**

- **esses desafios técnicos comprometem soluções tradicionais de remuneração por direitos autorais**, que dependem essencialmente da quantificação do uso das obras para estabelecer os pagamentos. Sem uma medição precisa, acordos de licenciamento podem acabar privilegiando grandes detentores de direitos com recursos mais jurídicos e prejudicando criadores independentes, que não conseguiriam medir uso de suas obras.

Questionados sobre os efeitos de limitações severas à disponibilidade de dados devido à potencial aplicação de regras de licenciamento e direitos autorais no Brasil, os especialistas destacaram que, em razão da natureza global da internet, **o treinamento de IAG poderia ser facilmente realizado em outros países, enfraquecendo o ecossistema de IA nacional e tornando as regras locais ineficazes**, com impactos negativos para a credibilidade regulatória e competitividade do país.

Outros impactos apontados foram:

- **Redução da qualidade dos modelos:** Modelos treinados com menos dados podem apresentar menor precisão e menor capacidade de generalização;

- **Aumento dos custos de desenvolvimento:** A necessidade de negociar licenciamento individual de cada dado encareceria o processo, tornando-o inviável para startups;

- **Concentração de mercado:** Empresas com acesso exclusivo a grandes *datasets* teriam vantagem competitiva, prejudicando a inovação aberta;

- **Efeitos Econômicos:** Caso o “uso” não seja o principal critério adotado, outras modalidades de mensuração podem gerar distorções de mercado, reforçando desigualdades estruturais no setor;

- **“Fuga de Centros de IA”:** Caso regras rígidas sejam implementadas no Brasil, a tendência seria a saída de centros de desenvolvimento de IAG do Brasil para outras jurisdições.

A principal contribuição deste estudo é demonstrar que a **compreensão técnica e realista do funcionamento dos modelos de IAG é vital para uma boa regulação**, e que é urgente ampliar a participação de profissionais STEM no processo legislativo.

SUMÁRIO

sumário interativo

1. Introdução

Imagine uma pessoa buscando informações sobre a economia do Brasil nos anos 1990. Em vez de acessar uma reportagem ou livro, ela pergunta a um *chatbot* de inteligência artificial. Em segundos, o sistema responde com um texto claro, bem estruturado e preciso. Embora nenhum conteúdo seja reproduzido literalmente, o modelo foi treinado com grandes volumes de textos disponíveis na internet, como matérias jornalísticas da época, trabalhos acadêmicos e verbetes da Wikipedia.

Será que a forma como esse *chatbot* utilizou os textos conflita com direitos dos autores e autoras dos textos que foram utilizados durante o treinamento? Essa é uma pergunta difícil, e que faz parte de um debate bastante amplo: **a complexa relação entre IAG e direitos autorais**.

Direitos autorais
são regras que protegem criadores e criadoras de obras intelectuais, como músicas, textos e imagens, permitindo que controlem o uso de suas criações e que possam receber remuneração quando outra pessoa usa essas obras. A Lei 9.610/1998 regula esses direitos no Brasil, incluindo suas exceções.

No entanto, essas discussões ocorrem frequentemente sem uma análise sobre aspectos técnicos da IAG – e é com esse recorte que propomos este estudo. Nossa ideia **é traduzir aspectos técnicos e operacionais da IAG para oferecer evidências que possam ser usadas no debate regulatório sobre IAG e direitos autorais**. Não buscamos hierarquizar dimensões políticas, econômicas ou jurídicas – afinal, é essencial que o debate seja orientado por múltiplas perspectivas. Ainda assim, acreditamos que a dimensão técnica – embora não seja a única relevante – é fundamental para que as discussões avancem com base em evidências concretas e soluções viáveis.

1.1. O que é Inteligência Artificial Generativa e por que isso importa?

Neste trabalho, consideramos tecnologias de Inteligência Artificial Generativa (IAG) como **sistemas que empregam técnicas estatísticas e de aprendizado de máquina (machine learning) para gerar novos textos, imagens ou outros tipos de conteúdo** (Daase et al, 2024). Diferentemente dos modelos analíticos que interpretam, classificam e tomam decisões baseadas em dados (Amorim, 2025), sistemas de IAG são capazes de gerar novos dados, como textos e imagens, utilizando padrões extraídos de extensas bases de dados de treinamento – os chamados *datasets*.

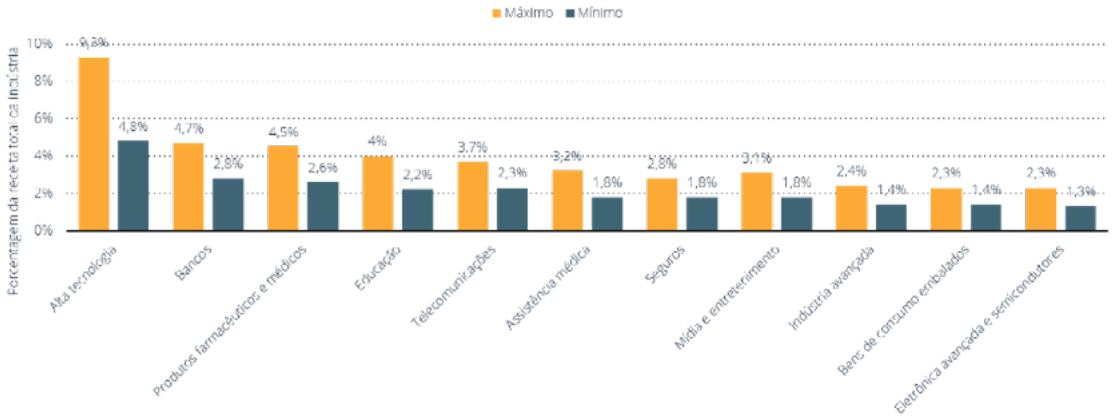
Datasets (ou **dados de treinamento**) são conjuntos organizados de dados – como textos, imagens ou vídeos – usados para treinar sistemas de IAG, e que ajudam a máquina a "*aprender*" padrões e melhorar suas respostas.

Figura 1. Diferenças entre modelos de IA Analítica e Generativa

Fonte: elaboração própria, a partir de Ramos (2023).

Sistemas baseados em IAG tiveram rápida adoção nos últimos três anos, e seu impacto econômico é significativo: estudo da McKinsey (Chui et al, 2023) projeta que o uso dessas tecnologias pode gerar até 5% de crescimento adicional para a economia global nos próximos cinco anos, afetando diretamente setores como agronegócio, seguros, bens de consumo e indústria farmacêutica.

Figura 2. Impacto econômico futuro da IA generativa em organizações no mundo em 2023, por setor econômico



Fonte: Chui et al (2023).

A economia da IAG não é composta por um ator único, mas por um ecossistema estruturado, com diferentes empresas exercendo funções complementares e em cadeia. Podemos resumir isso em **três camadas**[1]:

i. **infraestrutura**, composta por fabricantes de hardware responsáveis por chips e *data centers* de alto desempenho, essenciais para o processamento de grandes volumes de informação;

ii. **modelos**, composta por empresas que desenvolvem e licenciam modelos fundacionais, como os grandes modelos de linguagem (em inglês, *Large Language Models,* ou LLMs), baseados em redes neurais com bilhões de parâmetros treinados com vasta quantidade de dados e voltados para produção de texto; e

1 Outros estudos sugerem uma divisão diferente de camadas, em quatro ou seis (Simmons, A., 2023; Epical, 2024). Para fins didáticos, e baseando-se explicitamente no modelo de Benkler (2006), optamos por simplificar somente em três.

iii. **aplicações**, camada na qual empresas desenvolvem e oferecem sistemas de software que, a partir dos modelos e da infraestrutura, oferecem soluções e serviços para usuários finais — sendo os *chatbots* um dos exemplos mais conhecidos.

Quando falamos de **dados de treinamento, estamos preocupados especialmente com os modelos**. Mas é importante esclarecer: a forma como esses *datasets* são processados pelos modelos é bem diferente do que acontece em aplicações baseadas em armazenamento ou reprodução conteúdo – como um serviço de *streaming* de músicas ou séries, por exemplo. (veremos isso adiante, na apresentação dos resultados do estudo).

1.2. Mineração de Dados, IAG e Direitos Autorais

Há ao menos dois debates sobre direitos autorais e IAG que merecem ser diferenciados: um sobre a proteção de obras *criadas* por sistemas de IA (*quando a IA cria uma música, quem é o autor?)*, e o outro sobre obras protegidas por direitos autorais que estão presentes nos dados de treinamento dos modelos. **Aqui vamos falar exclusivamente desse segundo debate – e que começou bem antes da IAG** (Fill-Flynn et al, 2022).

Isso porque a prática de **mineração de dados** –processo que envolve métodos estatísticos para identificar padrões e correlações entre dados – começou a se popularizar ainda nos anos 1990 (Coenen, 2004). Foi nessa mesma época que surgiu a técnica de ***crawling*, ou varredura de dados,** em que um sistema percorre sites, páginas ou bancos de dados para analisar conteúdos, indexá-los e, em seguida, incorporá-los em práticas como a mineração de dados.

***Crawling*, mineração de dados e aprendizado de máquina são técnicas que servem como base não só para IAG, mas para aplicações tão diversas quanto buscadores de internet, ferramentas de comparação de preços, serviços de indexação de artigos científicos e plataformas que monitoram dados abertos de governos.**

RECAPITULANDO:

***Crawling*:** coleta automática de dados, como textos e imagens, para formar bancos que servirão de base para análise.

Mineração de dados: processo de analisar grandes volumes de dados para identificar padrões e correlações, antes ou independentemente do uso em treinamento de sistemas de IAG.

Treinamento de dados: etapa em que um sistema de IAG aprende com os dados, ajustando parâmetros para reconhecer melhor padrões e gerar respostas.

Aprendizado de máquina: processo pelo qual um sistema melhora continuamente suas respostas ao identificar padrões nos dados durante o treinamento.

A importância desses processos é tão grande que, nos últimos anos, diversos países passaram a incluir em suas leis de direitos autorais exceções para atividades de *data mining*, especialmente para pesquisa ou inovação no setor público e privado:

Na Europa, países como Reino Unido e Alemanha possuem exceções amplas para treinamento e mineração de dados, e recentemente a União Europeia, por meio do *AI Act,* também incorporou regras específicas sobre esse tema (Rosati, 2024);

O Japão tem se destacado por uma postura de incentivar o uso de dados para pesquisa e desenvolvimento nos setores público e privado (Ueno, 2025);

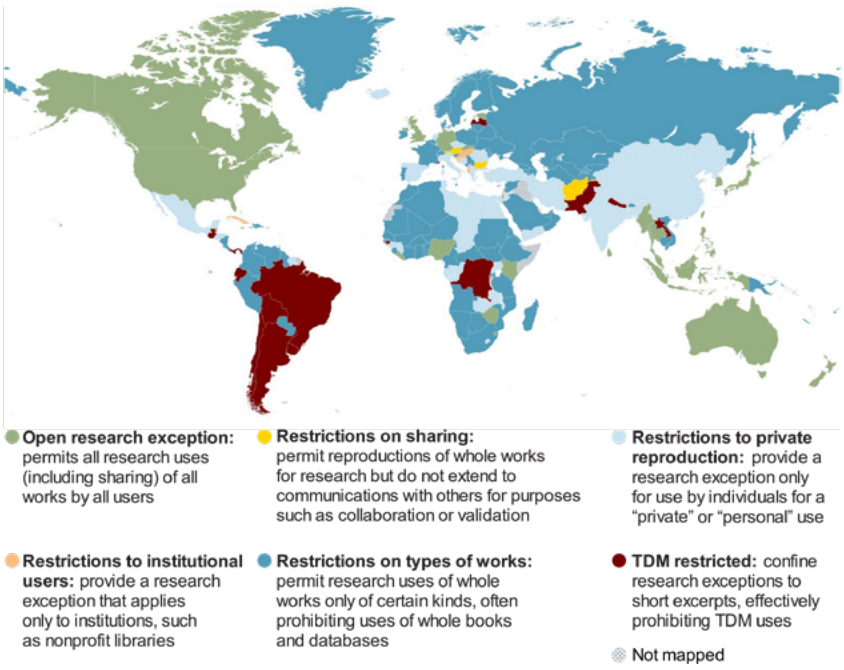
Nos Estados Unidos, a doutrina judicial do *fair use* tem sido interpretado como uma exceção válida para mineração e treinamento de dados; contudo, recentes discussões judiciais têm gerado insegurança jurídica sobre a interpretação desse conceito no âmbito da IAG, com críticas inclusive de importantes organizações da sociedade civil (Noble, 2025).

Já na China, há um cenário de insegurança judicial semelhante ao dos EUA, com a legislação sendo um pouco mais clara a favor da exceção de uso em comparação a estadunidense (Karaganis, 2024).

Na **América do Sul,** o cenário jurídico é bem diferente. Leis de direito autoral na região não criaram exceções específicas sobre treinamento e mineração de dados, um aspecto que gera insegurança jurídica para investimentos em *data centers* na região, além de impor barreiras ao desenvolvimento de tecnologias locais (Schirru et al, 2024).

A dependência de modelos treinados em outras jurisdições pode limitar a capacidade dos países latino-americanos de desenvolver tecnologias alinhadas às suas realidades culturais, linguísticas e sociais, e aplicações em áreas como saúde pública, justiça, educação ou cultura local podem ser especialmente afetadas.

Figura 3. Exceções de Direitos Autorais para Pesquisa, Treinamento e Mineração de Dados



Fonte: Fill-Flynn et al, 2022.

1.3. O Momento do Debate no Brasil: O PL 2.338/23

O debate legislativo sobre IAG no país ganhou novo impulso com a aprovação, pelo Senado Federal, do Projeto de Lei nº PL 2.338/23. O projeto originou-se de um anteprojeto elaborado por uma comissão de juristas e apresentado pelo presidente do Senado, Rodrigo Pacheco, em 2023, e incorpora dispositivos de outras sete propostas legislativas, incluindo o PL 21/2020 — já aprovado pela Câmara dos Deputados em 2021, mas que estava com tramitação paralisada no Senado[2].

Inspirado no *AI Act* da União Europeia e em referências normativas da área de proteção de dados pessoais, o projeto propõe um regime de obrigações baseado em risco aliado a um conjunto de garantias para pessoas afetadas por sistemas de IA. **Entre os direitos assegurados, estão o acesso à informação prévia sobre a interação com sistemas automatizados, o direito à privacidade e à proteção de dados pessoais e o direito à não discriminação.** Para sistemas classificados como de alto risco, o texto também prevê salvaguardas adicionais, como o direito à explicação, à contestação e à revisão humana de decisões automatizadas.

Em relação aos direitos autorais, o PL 2.338/23 adotou uma abordagem mais restritiva do que a proposta da União Europeia e de outros países. Em resumo, os artigos 62 a 65 do projeto:

- Criam exceções para mineração e treinamento de IA somente para instituições científicas, educacionais, museus, arquivos e bibliotecas, desde que sem fins comerciais e com acesso lícito;
- Estabelecem que desenvolvedores de IA devem cumprir obrigações de transparência, como a divulgação pública das bases de dados usadas no treinamento; e
- Criam mecanismos de remuneração, permitindo negociação coletiva ou direta com titulares de direitos autorais, considerando porte e impacto econômico.

A proposta gerou reações imediatas. De um lado, setores culturais e representantes de criadores destacaram o ineditismo da medida e seu compromisso em garantir proteção aos direitos autorais na era digital. De outro, surgiram preocupações quanto à viabilidade técnica das exigências estabelecidas e ao impacto potencial da proposta sobre a competitividade do Brasil no cenário global de desenvolvimento e inovação em IA.

Mecanismos de remuneração de direitos autorais asseguram pagamento pelo uso de obras criativas. Alguns exemplos incluem:

- **Música:** titulares recebem por execuções públicas ou digitais. No Brasil, o ECAD faz a arrecadação e distribuição para autores, intérpretes e editoras.
- **Audiovisual:** roteiristas e diretores são pagos pela exibição de obras.

2 Fonte: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262 . Acesso em: 12 mai. 2025.

"Em qualquer tipo de atividade econômica existe o insumo que é fundamental, e quem coordena aquela atividade tem de pagar por ele. No caso da inteligência artificial, o principal insumo é a criatividade, é o que cada um foi capaz de criar, e que vai ser minerado pela empresa que vai desenvolver o programa de inteligência artificial, que também terá de pagar por isso por conta da criatividade que as pessoas inserem na sua produção musical, literária, no que quer que seja"

Senador Humberto Costa (PT-PE)[3].

o PL 2338 não acompanha as tendências internacionais que buscam alcançar um equilíbrio entre a proteção dos direitos autorais e o desenvolvimento da IA. Países como Singapura e Japão permitem amplamente o treinamento de modelos e sistemas de IA. A União Europeia, por sua vez, adotou regras mais flexíveis que permitem análise computacional de obras disponíveis publicamente para viabilizar o treinamento de IA, reconhecendo a importância de fomentar a inovação nesse campo, ao mesmo tempo em que garante que os detentores de direitos possam indicar, por meios técnicos, que não permitem o treinamento em suas obras.

ABAG - Associação Brasileira do Agronegócio[4].

Figura 4. Tabela comparativa de legislações relativas ao treinamento de modelos de IA

COMO O PL 2338/23 SE COMPARA COM A LEGISLAÇÃO DE OUTROS PAÍSES?

Pergunta	EUA	UE	China	Japão	Brasil (PL 2.338/23)
Modelos podem ser treinados a partir de obras protegidas por direitos autorais disponíveis publicamente?	✓ **Sim,** se "transformativo" (doutrina do *fair use*)	✓ **Sim,** salvo se o titular informou seu *opt-out*	✓ **Sim,** há exceção legal para treinamento de dados, embora não seja expressa	✓ **Sim,** há exceção legal para treinamento de dados	✗ **Não,** o treinamento não é uma exceção legal expressa, se houver finalidade comercial/ lucrativa
Titulares de direitos autorais podem impedir que suas obras sejam usadas no treinamento de modelos?	✗ **Não,** salvo se provarem judicialmente que não é um *fair use*	✓ **Sim,** é possível o opt-out técnico (via metadados) ou por licenciamento	✗ **Não,** o direito de *opt-out* não é previsto na lei	✗ **Não,** o direito de *opt-out* não é previsto na lei	✓ **Sim** – sendo a regra *opt-in*, empresas precisam negociar o uso antes do treinamento
Titulares de direitos autorais podem exigir indenização de empresas que usam suas obras no treinamento de modelos?	? **Incerto** – depende de uma decisão judicial	! **Parcial –** somente se o opt-out for descumprido	? **Incerto** – há discussão judicial atual sobre o tema	! **Parcial –** a lei prevê o direito caso haja uso indevido ou plágio no *output*	! **Parcial –** o direito à indenização existe, mas sem clareza de critérios para seu cálculo

Fonte: elaboração própria.

3 Fonte: https://www12.senado.leg.br/noticias/materias/2024/12/10/senado-aprova-regulamentacao-da-inteligencia-artificial-texto-vai-a-camara. Acesso em: 12 mai. 2025.

4 Fonte: https://abag.com.br/regras-equilibradas-de-direitos-autorais-e-a-competitividade-do-brasil-em-inteligencia-artificial/. Acesso em: 12 mai. 2025.

1.4. A Proposta Metodológica dessa Pesquisa

Este contexto reforça a relevância dessa pesquisa: em um momento no qual o Brasil discute seu marco legal sobre IA, **formuladores de políticas públicas precisam compreender com profundidade as dimensões técnicas envolvidas**. Questões como a possibilidade de rastrear e quantificar o uso de conteúdos autorais, atribuir sua contribuição individual ao resultado de um modelo, estimar os custos de um eventual sistema de compensação e avaliar quem de fato se beneficiaria com essas medidas **precisam ser respondidas com base em evidências, antes que soluções legislativas sejam definidas**.

> **Esta pesquisa tem como objetivo entender de que forma o treinamento de modelos de IAG envolve o uso de conteúdo protegido por direitos autorais, e quais são os desafios técnicos relacionados à viabilidade de propostas de remuneração associadas a esse uso.**

Neste estudo, combinamos duas abordagens metodológicas. A primeira é a *evidence translation*, ainda pouco explorada na governança digital no Brasil, que visa **produzir evidências robustas e acessíveis para decisões públicas** (Ingold, 2025).

> Sempre que usamos quadros em rosa, gráficos ou exemplos destacados na diagramação, fazemos isso de forma consciente. Sabemos que corremos o risco de imprecisões técnicas, mas entendemos que, **dentro da lógica de traduzir evidências complexas em conhecimento aplicado, tornar o conteúdo mais claro e acessível é uma escolha metodológica necessária** — e um posicionamento que assumimos com transparência.

A segunda é a abordagem qualitativa. Em vez de revisões de literatura e *desk research* tradicional, conduzimos **entrevistas semiestruturadas** para captar percepções e experiências de um grupo **frequentemente ausente no debate regulatório: profissionais de STEM** (ciências exatas, tecnologia, engenharia e matemática).

Inspirados em estudos de recepção, buscamos entender como esses profissionais interpretam desafios técnicos sobre a relação entre IA e direitos autorais. Ao longo de um mês, realizamos oito entrevistas com *experts* no tema da pesquisa, focando em profissionais de nível sênior e com experiência e formação acadêmica no campo de STEM. As entrevistas seguiram roteiros pré-definidos e protocolos de confidencialidade, sendo suas transcrições e memoriais avaliados por meio do software Atlas.ti a partir da técnica de análise temática.

Figura 5. Tabela descritiva das pessoas entrevistadas

Entrevistado(a)	descrição
1	mulher, doutora e cientista de dados de empresa brasileira de grande porte no setor de software
2	homem, doutor, cientista de dados e professor universitário na área de tecnologia
3	homem, cientista de dados em empresa brasileira de grande porte no setor financeiro
4	homem, doutor, cientista de dados e professor universitário na área de tecnologia e administração
5	homem, engenheiro de software e executivo em startup brasileira
6	homem, mestre, engenheiro elétrico, arquiteto de soluções de IA em *big tech*
7 *	mulher, profissional de inteligência artificial em *big tech*
8 *	homem, consultor de *machine learning* em startup brasileira

* entrevistas preliminares

Fonte: elaboração própria.

A metodologia completa, com detalhes sobre os procedimentos adotados, está ao final do estudo.

2. Resultados Principais

2.1. Quantidade, qualidade e diversidade de dados de treinamento impactam a performance do modelo

Os entrevistados explicaram que modelos de IAG são altamente dependentes da qualidade, diversidade e quantidade dos dados, e que não há uma hierarquia necessária entre esses fatores – isso vai depender do objetivo de determinado modelo.

- A **qualidade dos dados** é central: conteúdos com erros, vieses ou informações incompletas comprometem as inferências do modelo e podem reproduzir distorções ou omissões;

- Da mesma forma, a **diversidade dos dados** — em idiomas, culturas, estilos e contextos — é essencial para garantir respostas inclusivas e generalizáveis; e

- A **quantidade de dados** também é relevante, especialmente por conta do modelo matemático adotado – ***"em modelos neurais, quanto mais dados, melhor"***[5] , disse uma pessoa entrevistada.

Contudo, nenhum desses fatores isolados são garantia de bom desempenho. Modelos treinados com grandes volumes de dados homogêneos, por exemplo, podem reproduzir vieses e apresentar limitações de aplicabilidade. Como disse uma pessoa entrevistada:

> *"Tem que ter cuidado, porque quantidade não é número de fotos. Não adianta passar um trilhão de imagens se são todas parecidas".*

Algumas pessoas entrevistadas mencionaram que modelos menores e mais específicos podem ser até mais eficazes para certas aplicações, além de mais econômicos. Essa afirmação alinha-se com experimentos acadêmicos recentes, que buscam criar conjuntos de *datasets* menores, mas cuja diversidade e qualidade dos dados compensa sua limitação em quantidade (Gao et al, 2020; Leffer, 2025).

Interessante notar que alguns entrevistados também destacaram que alguns dos modelos mais populares já *esgotaram* os dados públicos existentes na internet, obtidos por meio de *crawling*. Isso significa que o que vai diferenciá-los agora será **(i) a performance técnica dos modelos**, como maior capacidade de processamento, inovação no formato de cálculos, personalização, entre outros fatores, ou **(ii) a incorporação, em suas bases, de datasets que não podem ser capturados por *crawling*,** mas cuja qualidade pode ser um diferencial no modelo, o que explica por que diversas empresas estão buscando adquirir licenças para uso de acervos históricos de jornais, geralmente privados e não disponíveis publicamente na internet (Barcott, 2025).

5 Com o objetivo de preservar o anonimato e a confidencialidade dos participantes da pesquisa, foram realizadas modificações pontuais nas citações apresentadas neste estudo. Em determinadas circunstâncias, procedeu-se a adaptações linguísticas específicas para assegurar a intenção original dos entrevistados na transcrição textual. A preservação do registro discursivo foi mantida sempre que possível, respeitando os princípios metodológicos estabelecidos.

Nem todos os dados podem ser coletados por *crawling*. Isso porque muitos estão protegidos por barreiras técnicas (como *paywalls*), exigem login e senha para acessar, ou têm restrições legais, como dados pessoais sensíveis. Além disso, há conteúdos em formatos não acessíveis automaticamente, como arquivos offline ou coleções privadas. Isso limita o alcance do *crawling* e exige outras formas de acesso ou autorização, e que podem envolver acordos financeiros entre as empresas.

Esses achados técnicos trazem lições diretas para o debate sobre direitos autorais e IAG, alertando que modelos de remuneração baseados apenas no volume de obras usadas **podem não capturar o real impacto de cada contribuição no desempenho de um sistema**. Uma abordagem mais equilibrada precisaria considerar não só a quantidade, mas também a qualidade e a relevância contextual das obras no treinamento – um desafio enorme do ponto de vista técnico, como veremos adiante.

A inviabilidade técnica de medir a contribuição de obras em IAG

As pessoas entrevistadas explicaram que modelos de larga escala não funcionam por indexação direta dos dados (como em uma biblioteca), mas operam por meio de padrões estatísticos extraídos dos dados. Cada obra é fragmentada em palavras, que são transformadas em bilhões de representações vetoriais sem vínculos diretos com os arquivos de origem – e que sequer são armazenados.

Assim, a tentativa de identificar quanto uma obra individual contribuiu para um resultado específico é tecnicamente inviável.

Isso porque, durante o treinamento, um modelo analisa grandes volumes de dados em cada *dataset* para ajustar representações matemáticas (próprias de cada modelo), **sem necessariamente copiar ou armazenar os dados**. Em outras palavras: enquanto um app de músicas processa informações para reproduzir, sistemas de IAG processam informações para generalizar.

É por conta dessa generalização que um sistema de IAG treinado em milhões de imagens pode produzir uma nova composição visual que reproduz características comuns de obras do século XVI, sem replicar nenhuma obra específica, somente incorporando elementos recorrentes de diversas referências.

EXEMPLO: COMO A IA TRANSFORMA NÚMEROS EM CONTEÚDO NOVO?

Durante o treinamento, a IA analisa milhões de dados e transforma cada um deles em uma **representação matemática** – o que é também chamado de **vetor**. Esses vetores representam características do que foi aprendido. Vamos dar um exemplo e ver como uma palavra pode ser transformada em um vetor de centenas (às vezes, milhares) de números:

cachorro → [0,2, -1,3, 7,8, **-0,4**, ...]

Esses números não têm significado isolado. O que **importa é como eles se relacionam com outros vetores**. A palavra "quente" pode não ser relacionada a cachorros, mas o sistema pode atribuir alguma correlação – que vai ser identificada pela repetição de algum dos números do vetor:

quente→ [65,1, -1,12, 32,8, **-0,4**, ...]

Com o tempo, o modelo vai aprendendo cada vez mais correlações, ou seja, como os diferentes vetores podem se conectar entre si. É um processo intensivo – estamos falando de trilhões de vetores! –, e que exige muita capacidade de processamento de dados e cálculos de altíssima complexidade, e que são conhecidos como **redes neurais,** pela semelhança com o funcionamento do sistema nervoso de organismos vivos.

Agora, vamos imaginar que você digite, em um *chatbot*, o seguinte comando:

"Complete a seguinte frase: No almoço, pedi um cachorro ____"

A primeira coisa que o modelo de IAG fará é transformar esse comando em **vetores**, transformando as palavras em sequências numéricas. Em seguida, o modelo irá buscar correlações: **quais desses números se relacionam com outros números** que o modelo já conhece**. É um cálculo de probabilidade**: o modelo não escolhe palavras aleatoriamente, mas sim **a próxima palavra mais provável**. É como se o modelo se perguntasse: *"Com base nesse comando, qual é a palavra mais provável que venha agora?"*. Vamos simplificar e olhar novamente os nossos vetores de exemplo:

Almoço→ [0,2, -1,3, 7,8, **-0,4**, ...]

cachorro → [65,1, -1,12, 32,8, **-0,4**, ...]

quente→ [0,7, -8,3, 7,1, **-0,4**, ...]

O modelo parece ter encontrado uma correlação! Se essa for a mais provável estatisticamente, o modelo então responderá da seguinte forma:

No almoço, pedi um cachorro-quente.

O mais interessante é que, mesmo que o modelo tenha aprendido com milhares de frases e conteúdos, **esse resultado, por mais simples que pareça,** é **uma combinação nova de palavras**, gerada a partir de padrões estatísticos.

Essa característica distingue os modelos de IA de outras aplicações, como serviços de streaming – que funcionam mais como *bibliotecas digitais*. Nesses casos, o consumo pode ser vinculado a uma unidade de conteúdo, **sendo possível atribuir o uso ao resultado**. Já nas técnicas de IAG, não há estrutura de metadados ou sistema de rastreamento que permita reconstruir as relações de causa e efeito entre dados de entrada (*input*) e de saída (*output*).

Uma das pessoas entrevistadas explicou essa questão:

> *"Eu consigo afirmar que o autor X foi usado para treinar um modelo, mas ele não vai conseguir dar a resposta precisa: "para essa resposta, eu usei tal texto do autor X" (...) ou seja, ele usou na resposta o texto do Autor X, que estava no treinamento, ou pode ser, por exemplo, outros textos de pessoas que escreveram sobre Autor X?"*

Vamos imaginar que o modelo aprenda a correlacionar a palavra "chuva" com "tristeza", a partir de duas letras de música (**Música A** e **Música B**) e três livros (**Livro A**, **Livro B** e **Livro C**). Essa correlação é tão forte que vai gerar um vetor específico.

Quando o modelo criar a frase **"chuva é tristeza"**, será possível auditar e identificar que o vetor (23;0.4;18.4;80) foi usado – mas não será possível descobrir qual das 5 fontes de dados contribuiu para esse resultado, já que não houve armazenamento nem indexação, somente um *aprendizado de máquina*.

Ou seja, a tentativa de isolar a influência de um único ou conjunto restrito de informações torna-se complexa. A ideia de que seria possível calcular o "peso" de uma obra individual no desempenho de um modelo contraria o funcionamento estatístico de sistemas de aprendizado de máquina, que aprendem por meio de padrões difusos e recorrências probabilísticas. Outro entrevistado explicou essa questão a partir da forma como os modelos entendem diferentes idiomas – e transformam os conceitos em vetores:

> *"Em um modelo, quando você escreve uma pergunta em português ou em inglês, o modelo primeiro que vai fazer é tomar essa frase da pergunta e levar uma representação matemática que já é agnóstica da língua. Isso é uma coisa muito linda. Imagina você pegar cachorro e ele vai levar isso para um vetor matemático que significa cachorro em qualquer língua. Em todas as línguas você vai chegar na mesma coisa".*

A discussão sobre qualidade também tangencia a ideia de valor cultural vs. valor estatístico: Uma obra pode ter enorme valor cultural (ex.: um trecho de um clássico da literatura), mas, para o modelo de IA, sua contribuição estatística específica pode ser irrelevante – o que torna complexo o desenvolvimento de um sistema de remuneração por direitos autorais. Essa é a conclusão de um experimento recente feito por De La Rosa et al (2024), demonstrando que trabalhos de ficção não são tão determinantes na performance dos modelos.

> Vamos imaginar que um modelo foi treinado com dois conjuntos (*datasets*) diferentes: o primeiro tem livros do autor X, e o segundo traz artigos acadêmicos que analisam a obra de X.
>
> Agora, pense em alguém pedindo ao chatbot: "Escreva um parágrafo original no estilo do autor X." Sabemos que o modelo usou informações de ambos os conjuntos para gerar a resposta. Mas, como esses dados foram transformados em números e padrões estatísticos durante o treinamento, não há como dizer qual conjunto pesou mais na construção do texto final.
>
> Mesmo que ambos tenham ajudado, não conseguimos medir quem foi mais importante. E aí surge a pergunta: **como remunerar justamente quem mais contribuiu, se não dá para identificar o peso de cada parte?**

Por fim, alguns entrevistados ressaltaram que essa questão não parece ser uma escolha de mercado, mas sim uma limitação do estado da arte da tecnologia. Nesse sentido, nossas pesquisas mostraram um interesse recente em conferências e experimentos acadêmicos em universidades, como por exemplo os *working papers* de Wang et al (2024) e Zhang et al (2025) utilizam técnicas de teoria dos jogos para tentar estimar esses pesos, mas reconhecem uma série de limitações metodológicas, como a complexidade computacional, a fragmentação dos dados em diferentes fontes e identificação precisa de quais obras estão ou não protegidas por direitos autorais.

2.2. A redução de dados pode ocasionar redução na qualidade dos modelos de IAG

É unânime, entre as pessoas entrevistadas, que a restrição ao uso de dados — por regulações, custos ou riscos legais — impacta diretamente a qualidade dos modelos. **Quanto menor a base de dados disponível, mais limitado será o universo que o modelo poderá representar**, resultando em produtos que tendem a ser mais pobres em nuance, precisão e aplicabilidade.

Algumas pessoas entrevistadas comentaram sobre a substituição por dados sintéticos – criados artificialmente para suprir limitações de diversidade na base de dados –, mas foram categóricos em afirmar que:

> *"A precisão não fica tão boa como se você usasse os dados reais mesmo de fato".*

Uma questão que nos pareceu relevante nas entrevistas é que **a redução de dados em português pode ocasionar uma piora expressiva na qualidade dos modelos para questões de representação cultural local:** a língua portuguesa representa pouco menos de 4% do conteúdo aberto da internet, enquanto inglês representa quase metade (Statista, 2025) – ou seja, há um risco real de que os modelos se tornem menos relevantes para o público local. Como disse uma pessoa entrevistada:

> *"[Se o modelo] é treinado só com dados de outros países, porque o Brasil fecha, provavelmente esses modelos não vão conseguir trabalhar com problemas típicos brasileiros ou com algumas coisas que são específicas do Brasil. Então, se você perguntar quem é o campeão do campeonato brasileiro, ele não vai saber, porque ele não pode usar essa informação, a menos que isso tenha sido publicado em alguma outra fonte externa que ele não tá usando".*

2.3. O licenciamento individual de dados pode tornar o desenvolvimento de modelos brasileiros inviável

Perguntados sobre os impactos de um sistema regulatório que exigisse o licenciamento obrigatório de obras para o treinamento de modelos de IAG, **as pessoas entrevistadas concordaram que os impactos seriam graves, especialmente para as empresas brasileiras**. Como colocou uma entrevistada:

> *"As maiores prejudicadas seriam as empresas brasileiras mesmo, porque pra gente ter uma operação no Brasil e fazer esse treinamento no Brasil, a gente não conseguiria, ficaria inviável".*

A preocupação que surgiu nas entrevistas foi tanto em relação à operacionalização, quanto o custo desses licenciamentos. Uma das entrevistadas afirmou que a diversidade de criadores e fontes na internet é tão grande que seria praticamente impossível licenciar individualmente todos os conteúdos:

> *"O problema é se eu faço uma lei do tipo, uma lei geral, que eu quero que todo mundo se compense, sabe? Eu duvido até como vou compensar todo mundo. Eu estou achando uma foto lá na internet, outra aqui, outra aqui, que está publicamente disponível e como vou compensar cada uma dessas pessoas? Tipo, acho que o problema seria isso, seria fazer uma lei que na prática seja impraticável".*

O impacto financeiro também foi apontado como uma limitação relevante para o surgimento de novas empresas brasileiras, o que tornaria o desenvolvimento ***"impagável para startups"***. Esse cenário compromete o dinamismo do ecossistema de inovação, e cria também barreiras de competitividade do país no cenário internacional, já que outros países (como vimos na Introdução) estão justamente buscando formas de flexibilizar o uso de dados de treinamento. ***"Seria um problema que pode deixar um país pra trás"***, como disse um dos entrevistados.

2.4. Concentração de mercado: o acesso exclusivo a datasets pode beneficiar somente grandes empresas

Em um contexto de regulamentação rígida, **empresas com grandes bases proprietárias ou acesso exclusivo a dados podem ocupar posições ainda mais dominantes no mercado de IAG**. Segundo as pessoas entrevistadas, isso pode ser pode criar distorções de ambos os lados – entre os detentores de conteúdo e as desenvolvedoras de modelos de IAG.

Como mostra Barcott (2025), grandes empresas de modelos de IAG já estão desenvolvendo acordos exclusivos com empresas que possuem grandes bases de dados proprietárias. A questão, trazida em diversas entrevistas, é que **a diversidade de conteúdo na internet, combinada com a dificuldade de atribuição técnica, tornaria praticamente impossível a compensação financeira individualmente a pequenos criadores** – ainda que, estatisticamente, até mais relevantes do que as de grandes bases de dados.

> *"Nós não estamos falando só de nichos como as empresas de mídia, nichos como publishers de livros. Sinto que realmente começa a ser uma história em que quase que todo o website, na internet agora, tem direito de cobrar copyrights se eles assumem, a ser compensado por copyright, se eles assumem que o modelo foi treinado com isso".*

Do lado do desenvolvimento das tecnologias de IAG, essa concentração já é visível internacionalmente, com poucas empresas dominando o desenvolvimento de modelos, e que teriam mais condições de pagar altos custos relacionados com o licenciamento – além dos custos próprios do treinamento em si, que já são altíssimos, como destacou uma pessoa entrevistada:

> *"Você pode pôr o treinamento, mas tem que pôr os dinheiros também, tem que pôr os PHDs trabalhando nisso, enfim, tem um monte de custo indireto".*

2.5. "Fuga de centros": uma regra rígida no Brasil seria facilmente contornada – com efeitos econômicos e sociais relevantes

A imposição de regras excessivamente restritivas sobre o uso de dados para treinamento de IA pode ter como efeito colateral o que chamamos **de "fuga de centros" — o deslocamento de polos de inovação e investimento para países com regulações mais flexíveis**.

As pessoas entrevistadas comentaram que empresas de IAG poderiam, tecnicamente, deslocar suas atividades para jurisdições onde obrigações de remuneração por direitos autorais não existam. **Considerando a característica global e aberta da internet, isso seria muito simples de executar – e igualmente simples de contornar.** Como citou uma pessoa entrevistada:

> *“É como se você fosse proibir aqui, mas você não vai proibir no resto do mundo (...) então você decide, você não quer ter a tecnologia aqui e todos os seus vizinhos terem?”.*

Essa situação iria de encontro a políticas de incentivo a datacenters locais, e daria uma vantagem competitiva para empresas maiores, com maior infraestrutura em nuvem distribuída globalmente e que podem optar por treinar modelos em servidores localizados onde a lei seja mais favorável.

> *“Com certeza. Sem nenhuma dúvida. Pensa o seguinte: primeiro onde o treinamento acontece? Você tem diferentes escalas (...) [as empresas hoje] estão treinando em clouds que existem em vários países. E tem data centers em vários países. Então, se a pergunta é, hoje a tecnologia para treinar já está em vários países, com certeza, não tem nenhuma dúvida que está e vai estar mais ainda”.*

Além disso, se uma lei for apenas local, **sua eficácia sobre uma entidade estrangeira que disponibiliza um modelo via internet é limitada**. A menos que haja bloqueios totais de sites e aplicações (uma medida grave), a empresa estrangeira poderia oferecer seu serviço aos brasileiros de qualquer forma – o que poderia afetar a credibilidade da regulação no país.

> *“E o que eu posso fazer também é o seguinte: se eu tenho esse cuidado nos Estados Unidos e eu não tenho no Brasil. Por causa dessa restrição, eu posso mandar meu modelo para os Estados Unidos, treinar a parte que eu não tenho conhecimento lá, trazer ele de volta e continuar o treino no Brasil”.*

3. Análise e Comentários

Esta seção analisa os resultados da pesquisa, relacionando-os com literatura acadêmica e opiniões especializadas, por meio das lentes do autor e autoras deste trabalho.

3.1. O déficit de expertise técnica na formulação de políticas públicas sobre IAG

Após a realização das entrevistas e um reexame do texto sobre direitos autorais aprovado no PL 2338/23, pareceu-nos haver uma distância enorme entre a proposta do projeto e sua viabilidade técnica. **O que pode ter motivado isso?**

É uma questão difícil de deduzir empiricamente. Contudo, nossa hipótese exploratória é que **o debate legislativo sobre IA e direitos autorais no Brasil foi conduzido sem uma compreensão aprofundada da tecnologia.** Há alguns fatores que reforçam esse argumento.

Primeiro, o tema dos direitos autorais não figurou entre os eixos mais debatidos durante os trabalhos da Comissão do Senado. A análise das notas taquigráficas das 24 sessões da Comissão Temporária Interna sobre Inteligência Artificial no Brasil do Senado Federal (CTIA) mostrou que os debates se concentraram majoritariamente em tópicos como proteção de dados pessoais, classificação de riscos, definição de sistemas e impactos sobre a inovação. Embora os direitos autorais sejam uma dimensão relevante da regulação de IA, sua discussão foi significativamente menor em comparação com outros temas.

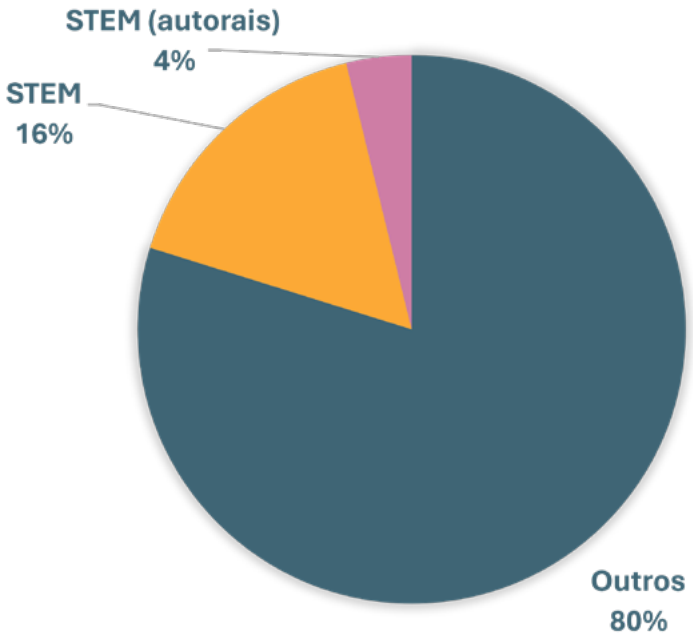
Figura 6. Nuvem de palavras das sessões da CTIA, a partir do uso do software Atlas.ti (ferramenta "Concepts").

Fonte: nuvem de palavras gerada por meio do software Wordclouds.com, a partir de informações geradas automaticamente pelo Atlas.ti, utilizando a ferramenta "Concepts"[6].

6 A partir do mapeamento automatizado, foi realizado um agrupamento temático de conceitos, aproximando expressões semelhantes (e.g. legislação e regulação; privacidade e dados pessoais). O tamanho de cada palavra reflete o cálculo ponderado da frequência dos conceitos após o agrupamento.

Além disso, observamos a **baixa presença de profissionais das áreas de STEM no debate** – ainda menor quando analisamos quantos deles discutiram, em suas falas, a questão dos direitos autorais a partir de sua visão técnica:

Figura 7. Gráfico perfil de participantes e contribuições da CTIA



Fonte: elaboração própria.

Ou seja, parece-nos que a ausência de especialistas técnicos na CTIA pode ter levado à proposição de **medidas que não dialogam com a realidade do funcionamento dos modelos de IA na questão de direitos autorais** – e parece-nos que essa desconexão entre regulação e tecnologia não é um problema isolado do Brasil, mas um desafio global. No entanto, para evitar a criação de leis que sejam inaplicáveis ou que prejudiquem a competitividade do país, é fundamental institucionalizar mecanismos de consulta técnica qualificada e baseada em evidências científicas e econômicas concretas, garantindo diretrizes factíveis.

3.2. Distorções econômicas: critérios além do "uso" favorecem empresas com grandes bases proprietárias, e não há evidência de como essa remuneração poderia chegar nos criadores

Os desafios técnicos na atribuição do uso de obras em sistemas de IAG comprometem a lógica econômica dos direitos autorais, que se baseiam na quantificação de como o conteúdo protegido é reproduzido, distribuído ou transformado para alocar sua remuneração (Watt, 2009). No entanto, essa justificativa - recompensar criadores proporcionalmente ao uso de sua obra - desmorona quando observamos que sistemas de IAG não conseguem mensurar rastrear confiavelmente o uso dessas obras.

Essa ruptura tem o potencial de distorcer os incentivos: estruturas de licenciamento que não sejam baseadas no uso podem agravar a concentração do mercado, e grandes detentores de direitos com recursos jurídicos podem negociar acordos de licenciamento em massa, enquanto criadores independentes, sem poder de barganha para comprovar o uso pela IAG, podem ser prejudicados, marginalizando vozes menores e reduzindo a diversidade criativa (Martens, 2024).

> **No atual estado da tecnologia de aprendizado de máquina, direitos autorais correm o risco de se tornar um instrumento que prejudica excessivamente a inovação em IA e também protegem inadequadamente criadores humanos.**

Abordar esses desafios em futuros estudos pode exigir uma **reconceitualização tanto do conceito de direito autoral, quanto do impacto da IAG nas indústrias criativas** através de uma perspectiva histórica mais ampla. Disrupções tecnológicas anteriores, como a transição da distribuição física para a digital, inicialmente geraram controvérsias, mas acabaram levando à transformação da indústria em vez do declínio, fomentando novos modelos de negócios e fluxos de receita que se aproveitaram de novas tecnologias para criar valor adicional (Masnick e Beadon, 2024).

4. Conclusão

O avanço da IAG levanta questões legítimas sobre como garantir um ecossistema que equilibre inovação e bem-estar social, e a regulação surge como uma ferramenta para promover esse equilíbrio. No entanto, os achados deste estudo sugerem que, para que essa regulação seja eficaz, suas diretrizes **precisam ser tecnicamente viáveis**.

Este estudo mostra que, embora seja possível identificar as bases de dados usadas no treinamento de modelos, ainda **não há soluções escaláveis e confiáveis para medir a contribuição específica de cada obra** em modelos de larga escala. Atualmente, isso se mostra como uma limitação **estrutural da tecnologia**, especialmente no aprendizado de máquina.

Logo, propostas regulatórias que não levem isso em conta pode gerar mensurações arbitrárias. Além disso, as entrevistas destacaram que restrições **excessivas ao uso de dados podem** criar barreiras de entrada para startups, pesquisadores independentes e instituições públicas, favorecendo a concentração de mercado. Também deve ser considerada a possibilidade de que empresas realoquem seus processos de treinamento para países com regulações mais permissivas, o que afeta a **credibilidade regulatória e a competitividade do país**.

Os achados desta pesquisa não devem ser usados isoladamente contra **a regulação**, ou contra a valorização dos direitos de criadores e criadoras. Pelo contrário, suas inferências apontam para a necessidade de uma regulação baseada **em evidências, que considere a realidade do setor e os limites técnicos da tecnologia existente**.

Como mensagem final, destaca-se a urgência de também ampliar a participação de especialistas técnicos no processo regulatório da IA. A atual tramitação legislativa demanda um diálogo efetivo com a comunidade técnica - não para sobrepor sua visão às demais, mas para garantir que as políticas públicas reflitam a complexidade dos sistemas a serem regulados.

4.1. Sugestões Para Futuros Estudos

Este estudo analisou a viabilidade de sistemas de remuneração por direitos autorais na IA, mas algumas questões permanecem abertas. Abaixo, listamos eixos de pesquisa que podem aprofundar o debate e subsidiar políticas públicas.

- **Impacto Econômico da IAG:** há poucas evidências sobre se a IAG gera prejuízos ou novas oportunidades para criadores. Pesquisas podem analisar como diferentes setores são impactados, avaliar mudanças na distribuição de renda e testar alternativas de monetização.

- **Percepção dos Criadores sobre o Uso de Seus Dados no Treinamento de IA:** a regulação muitas vezes ignora as percepções dos criadores. Pesquisas qualitativas com o olhar da recepção midiática podem explorar como criadores avaliam o uso de seus dados, seu nível de aceitação ou rejeição e sua visão sobre o debate regulatório.

- **Dinâmicas de Interesse no Debate Regulatório:** estudos podem mapear quem são os principais participantes no processo legislativo, como influenciam políticas, quais suas agendas de interesse e se há equilíbrio na representatividade dos diferentes setores.

- **Modelos de Remuneração: Existe Caminho Viável?:** Diante da falta de rastreabilidade, pesquisas podem avaliar os impactos de modelos de compensação estimada, *opt-out,* ou exceções ao direito autoral, a partir de metodologias econométricas ou avaliações de custos e benefícios sociais.

- **O Efeito de Restrições de Dados na Inovação e Competitividade:** Estudos podem medir como restrições impactam a qualidade dos modelos de IA, se favorecem grandes players e como incentivam a migração de empresas para jurisdições mais flexíveis.

Referências

AMORIM, P. **Analytical AI: A Better Way to Identify the Right AI Projects.** Disponível em: https://sloanreview.mit.edu/article/analytical-ai-a-better-way-to-identify-the-right-ai-projects/. Acesso em: 10 mai. 2025.

AUDENHOVE, L. V.; DONDERS, K. **Talking to People III: Expert Interviews and Elite Interviews.** In: VAN DEN BULCK, H.; PUPPIS, M.; DONDERS, K.; VAN AUDENHOVE, L. (Eds.). The Palgrave Handbook of Methods for Media Policy Research. Palgrave Macmillan, 2019.

BARCOTT, B. **How the Emerging Market for AI Training Data is Eroding Big Tech's "Fair Use" Copyright Defense,** 2025. Disponível em: <https://www.techpolicy.press/how-the-emerging-market-for-ai-training-data-is-eroding-big-techs-fair-use-copyright-defense>. Acesso em: 12 mai. 2025.

CHUI, M. et al. **Economic potential of generative AI.** McKinsey, 2023. Disponível em: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier. Acesso em: 11 mai. 2025.

COENEN, F. **Data Mining: Past, Present and Future**. The Knowledge Engineering Review, v. 00, p. 0–1, 2004.

DE LA ROSA, J., et al. **The Impact of Copyrighted Material on Large Language Models: A Norwegian Perspective.** arXiv preprint, 2024. Disponível em: <https://arxiv.org/html/2412.09460v1>. Acesso em: 11 mai. 2025.

DAASE, C. et al. **On the Current State of Generative Artificial Intelligence: A Conceptual Model of Potentials and Challenges**. 26th International Conference on Enterprise Information Systems, 2024.

FIIL-FLYNN, Sean M. et al. **Legal reform to enhance global text and data mining research.** *Science*, v. 378, p. 951-953, 2022.

GAO, L. et al. **The Pile: An 800GB Dataset of Diverse Text for Language Modeling**. arXiv preprint, 2020, disponível em: < https://arxiv.org/abs/2101.00027>. Acesso em: 10 mai. 2025.

GUEST, G.; BUNCE, A.; JOHNSON, L. **How Many Interviews Are Enough? An Experiment with Data Saturation and Variability**. Field Methods, 18(1), 59-82, 2006.

HERZOG, C; HANDKE, C.; HITTERS, E. **Analyzing Talk and Text II: Thematic Analysis.** In: VAN DEN BULCK, H.; PUPPIS, M.; DONDERS, K.; VAN AUDENHOVE, L. (Eds.). The Palgrave Handbook of Methods for Media Policy Research. Palgrave Macmillan, 2019.

INGOLD, Jo; MONAGHAN, Mark. **Evidence translation: an exploration of policy makers' use of evidence**. Policy & Politics, v. 44, n. 2, p. 171-190, 2016.

KARAGANIS, J. **Emerging Copyright Governance Frameworks Across the US, China, and Europe**. AI, Media & Democracy, 2024. Disponível em: < https://www.aim4dem.nl/is-ai-training-infringement/>. Acesso em: 12 mai. 2025.

LEFFER, L. **When It Comes to AI Models, Bigger Isn't Always Better,** 2025. Disponível em: <https://www.scientificamerican.com/article/when-it-comes-to-ai-models-bigger-isnt-always-better/>. Acesso em: 12 mai. 2025.

MARTENS, Bertin. **Economic arguments in favour of reducing copyright protection for generative AI inputs and outputs.** Working Paper, Bruegel, 2024.

MASNICK, M.; BEADON, L. **The Sky Is Rising: A detailed look at the state of the entertainment industries, 2024 Edition.** Copia Institute & CCIA Research Center, 2024.

NOBLE, T. **AI and Copyright: Expanding Copyright Hurts Everyone—Here's What to Do Instead**. Electronic Frontier Foundation, 2025.

RAMOS, P. H. (coord). **Governança Digital em Foco: estratégias para uso de IA generativa nas empresas**. Gtech – Grupo de Estudos em Direito e Tecnologia. Relatório de Pesquisa. São Paulo: Ibmec SP, 2023.

ROSATI, E. **Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law**. European Journal of Risk Regulation, p. 1–25, 31 out. 2024.

SALDAÑA, Johnny. **The Coding Manual for Qualitative Researchers**. 4. ed. Thousand Oaks: SAGE Publications, 2021

STATISTA SEARCH DEPARTMENT. **Languages most frequently used for web content,** 2025. Disponível em: < https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/>. Acesso em: 12 mai. 2025.

SCHIRRU, L. et al. Text and Data Mining Exceptions in Latin America. **IIC - International Review of Intellectual Property and Competition Law**, 19 set. 2024.

SOBEL, B. L. W. Artificial intelligence's fair use crisis. **Columbia Journal of Law & the Arts**, v. 41, p. 45-96, 2017.

UENO, H. **Japan's New Approach to Collaborative International R&D**. Issues in Science and Technology. Vol. XLI, Winter, 2025.

WANG, J. T. et al. **An Economic Solution to Copyright Challenges of Generative AI**. arXiv preprint, 2024. Disponível em: <https://arxiv.org/abs/2404.13964>. Acesso em: 12 mai. 2025.

Zhang, L. et al. **Fairshare Data Pricing for Large Language Models.** arXiv preprint, 2025. Disponível em: <https://arxiv.org/html/2502.00198v1>. Acesso em: 12 mai. 2025.

Anexo de Metodologia Reglab

FORMATO: POLICY BRIEF

Título	Remuneração por Direitos Autorais em IA: Limites e Desafios de Implementação
Pergunta de Pesquisa	**de que forma o treinamento de modelos de IAG envolve o uso de conteúdo protegido por direitos autorais, e quais os desafios técnicos à viabilidade de propostas de remuneração associadas a esse uso?**
Resumo de Metodologia	Esta pesquisa adota uma abordagem **qualitativa**, combinando coleta de dados primários através de **entrevistas em profundidade com especialistas** (*expert interviews*) e análise de dados secundários (documentos, literatura e casos práticos). A escolha metodológica se baseia na natureza exploratória do tema: por ser um assunto emergente, com poucas experiências concretizadas de remuneração por dados de IA, torna-se valioso captar as **percepções, opiniões e conhecimentos** de especialistas.
Coleta de Dados	A coleta de dados utilizou a metodologia de **expert interviews** (Audenhove e Donders, 2019), com entrevistas qualitativas semiestruturadas, de caráter exploratório. A escolha por esse método se justifica pelo caráter técnico do tema e pela ausência de dados sistematizados sobre o problema investigado, tornando fundamental o conhecimento acumulado de especialistas atuantes na área. A amostra foi definida com base em critérios de diversidade e representatividade, incluindo: participação mínima de mulheres; presença de representantes da academia ou centros de pesquisa; profissionais de empresas brasileiras; e especialistas de grandes empresas de tecnologia. A seleção combinou amostragem por conveniência e técnica de *snowballing*. Foram contatadas 16 pessoas, das quais oito aceitaram participar da pesquisa; as demais declinaram por indisponibilidade. As entrevistas foram realizadas entre os dias 12 e 31 de março de 2025, em formato online (via Teams), com duração média entre 45 e 60 minutos. Cada sessão contou com a presença de, no mínimo, dois pesquisadores do Reglab. O roteiro de perguntas utilizado encontra-se anexo. Dentre as entrevistas, duas foram conduzidas em caráter preliminar, com o objetivo de testar a estrutura do roteiro e validar hipóteses iniciais. Essas entrevistas não foram incluídas no processo de codificação, mas contribuíram substancialmente para o delineamento final da coleta. As seis entrevistas analisadas foram consideradas suficientes para fins de saturação teórica, dado que, em abordagens qualitativas com entrevistas semiestruturadas e em profundidade, a recorrência temática e a densidade analítica tendem a se consolidar com um número reduzido de participantes (Guest et al, 2006). Todas as entrevistas foram gravadas com autorização dos participantes, transcritas integralmente e acompanhadas por memorandos das pessoas entrevistadoras. O material foi armazenado e codificado no software Atlas.ti. Os nomes e instituições dos entrevistados foram anonimizados.
Análise de Dados	Os dados foram analisados por meio de análise temática, conforme Herzog et al. (2019), com dois ciclos de codificação indutiva. O primeiro ciclo consistiu em codificação conceitual aberta, e o segundo utilizou *pattern coding* para o agrupamento e refinamento das categorias analíticas (Saldaña, 2021). O processo foi realizado no software Atlas.ti. A escolha pela análise temática se justifica por sua adequação a estudos exploratórios que buscam estruturar e interpretar informações técnicas, permitindo a identificação de padrões conceituais em contextos de alta complexidade. A equipe adotou uma postura reflexiva ao longo do processo analítico, com registro de memorandos interpretativos e discussão sistemática de potenciais vieses analíticos. Os temas foram definidos com base na recorrência, densidade conceitual e relevância para os objetivos da pesquisa. As categorias finais incluíram, entre outras: “impossibilidade técnica de atribuição”, “modelos de remuneração”, “concentração de mercado”, “limites da rastreabilidade” e “impacto regulatório”. Para apoiar a análise crítica e a triangulação de evidências, foram utilizados os recursos de visualização, mapeamento e correlação do Atlas.ti. A análise foi conduzida entre os dias 2 e 15 de abril de 2025.

Procedimentos de Redução de Vieses	**Referências teórico-metodológicas consolidadas:** as técnicas de coleta e análise de dados adotadas neste estudo seguiram práticas reconhecidas na literatura acadêmica. A abordagem metodológica foi discutida internamente antes e após a realização das entrevistas preliminares, permitindo a incorporação de críticas e sugestões ao desenho final da pesquisa, antes do início do processo de análise **Categorização aberta:** a codificação dos dados seguiu uma lógica indutiva, sem categorias pré-definidas, permitindo que os códigos e temas emergissem diretamente do material empírico. Essa escolha metodológica visou minimizar vieses interpretativos decorrentes de imposições conceituais anteriores. **Triangulação de métodos:** os achados empíricos foram contrastados com análise documental de fontes secundárias, com o objetivo de comparar, validar e reforçar a consistência das interpretações construídas a partir das entrevistas. Essas referências foram expressamente citadas ao longo do texto. **Dupla validação em etapas críticas:** a codificação foi conduzida e revisada por dois pesquisadores de forma cruzada. A definição final dos temas foi realizada em discussão coletiva entre os três autores, assegurando múltiplas perspectivas e controle de vieses individuais na interpretação dos dados. **Registro e transparência metodológica:** todas as etapas do processo analítico foram documentadas, incluindo versões sucessivas dos arquivos e decisões de codificação. Essa prática permite a rastreabilidade do percurso metodológico, conforme as diretrizes do Reglab para transparência e replicabilidade.
Outras Limitações Metodológicas	**Escopo qualitativo e não generalizável:** o número reduzido de entrevistas priorizou profundidade analítica, mas não permite inferências estatísticas. **Amostragem por conveniência e redes de contato:** a seleção pode ter refletido vieses de disponibilidade e círculos profissionais, apesar dos critérios de diversidade. **Evolução tecnológica e regulatória:** os achados refletem o estado da arte até o momento da pesquisa e podem ser impactados por mudanças futuras no setor. **Dependência de Ferramentas Externas:** ainda que seja um das ferramentas analíticas mais consolidadas no setor acadêmico, a análise dependeu significativamente do uso do software Atlas.ti.

Uso de Software

SOFTWARE	USO NA PESQUISA
Suíte MS Office	edição de texto, planilhas e gráficos, entrevistas (Teams)
Suíte Adobe C	diagramação e finalização de gráficos e ilustrações.
Atlas.ti	organização, codificação e análise dos dados qualitativos
Cockatoo	Transcrição de áudio das entrevistas em texto.
ChatGPT 4o	brainstorm, sistematização de informações, revisão gramatical (ortografia, gramática busca de sinônimos), adequação da linguagem, adequação ao Manual de Redação Reglab.
Notion AI	edição e revisão de texto (ortografia e gramática, busca de sinônimos, adequação de linguagem, traduções); organização da pesquisa, estruturação de cronograma.
Wordclouds	elaboração de nuvens de palavras
Lex.page	revisão avançada de texto (brevidade, clichês, legibilidade, voz passiva, afirmações sem evidências, repetições).

Diretrizes Éticas	**Financiamento da Pesquisa.** Esta publicação faz parte de uma série de publicações patrocinadas pelas empresas Google, meta e B/Luz, em que o RegLab mantém controle editorial das publicações. Diferentemente de pesquisas comissionadas, o RegLab determinou o escopo, objetivos e a metodologia desse estudo com completa autonomia. Os autores mantêm total independência profissional e responsabilidade pelo conteúdo e conclusões deste trabalho. **Tratamento de Dados Pessoais.** A pesquisa envolveu o tratamento de dados pessoais exclusivamente nas etapas de coleta e análise, de forma limitada e proporcional aos objetivos do estudo, de acordo com a Lei nº 13.709/2018 (LGPD). **Base legal:** todos os participantes autorizaram formalmente sua participação mediante assinatura de termo de consentimento, com ciência sobre os objetivos da pesquisa e uso dos dados; **Finalidade e adequação:** os dados foram utilizados exclusivamente para os fins desta pesquisa, compatíveis com o consentimento obtido, não sendo empregados para outras finalidades; **Minimização e anonimização:** informações pessoalmente identificáveis que não eram relevantes para os objetivos da pesquisa foram anonimizadas nas transcrições e excluídas da base ativa; **Sigilo e confidencialidade:** na apresentação dos resultados, os dados foram mantidos sob sigilo, e as citações foram ajustadas, quando necessário, para garantir a confidencialidade das fontes. Somente um número limitado de pesquisadores diretamente envolvidos no projeto teve acesso aos dados pessoais e documentos originais; **Registro e segurança da informação:** os arquivos foram armazenados com controle de acesso por senha e conforme políticas internas de segurança da informação do Reglab; **Retenção e descarte:** os dados serão armazenados por até 12 meses, exclusivamente para fins de auditoria metodológica e eventual replicação, sendo posteriormente eliminados; **Uso Responsável de Dados Públicos:** Embora alguns dos dados analisados sejam públicos, seu uso foi feito de maneira responsável e ética, com o objetivo exclusivo de pesquisa independente. **Transparência Metodológica:** A metodologia de pesquisa foi detalhada para garantir transparência e replicabilidade, contribuindo para a integridade científica e permitindo a validação independente dos resultados. **Não-discriminação e Respeito à Diversidade:** A pesquisa foi conduzida de maneira a respeitar a diversidade e evitar qualquer forma de discriminação.

ANEXO II - ROTEIRO SEMIESTRUTURADO DAS ENTREVISTAS

#	PERGUNTA
0	Na redação do produto de pesquisa, como você gostaria de ser chamada/chamada? Qual título podemos usar?
1	Para começar, pode contar um pouco sobre sua experiência profissional e sua atuação na área de IA/Machine Learning?
2	Como ocorre a seleção e o uso de dados para treinamento de modelos de IA?
3	Quais tipos de dados são mais críticos para o desempenho do modelo? Há características específicas que tornam um conjunto de dados mais valioso?
4	Se houvesse uma limitação na disponibilidade de dados devido a regras de licenciamento, quais seriam os impactos técnicos e práticos no desenvolvimento e no desempenho dos modelos?
5	Pensando na possibilidade de dados serem restritos/indisponíveis por conta de restrições autorais, qual/quais seriam os impactos para modelos de IA?
6	Na sua experiência, é possível rastrear e documentar quais dados específicos foram utilizados no treinamento de um modelo de IA?
7	Como isso é feito na prática?
8	Considerando a forma como modelos de IA são treinados, como seria possível identificar e remunerar cada obra utilizada no processo? Quais chances e desafios há nessa atividade?
9	Na necessidade de haver conteúdos protegidos, há técnicas para garantir que modelos de IA não os memorizem e os reproduzam? Pode explicar um pouco essa questão?
10	Se um sistema de remuneração obrigatória por direitos autorais fosse implementado, quais seriam os impactos para empresas e startups que desenvolvem IA?
11	Caso restrições ao uso de dados para IA fossem implementadas no Brasil, o treinamento dos modelos poderia simplesmente ser transferido para outros países? Isso já acontece em outros contextos?
12	Se houvesse um sistema ideal para equilibrar o uso de dados para IA e a proteção de direitos autorais, como ele funcionaria na sua visão? Existem soluções técnicas viáveis para esse problema?
13	Há algo que não perguntamos, mas que você considera importante sobre o uso de dados no treinamento de IA?