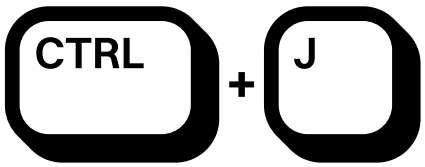


THE PROTOCOL GAP: BRASIL

Março 2026

Momentum
Journalism & Tech Task Force

International Fund
for Public Interest Media

Journalism
Relay
Project

Este relatório é uma publicação conjunta do Journalism Relay Project, Momentum – Journalism and Tech Task Force, e do International Fund for Public Interest Media (IFPIM).

Sérgio Spagnuolo (Journalism Relay Project) desenvolveu a metodologia e liderou a análise de dados. Ester Borges, Bruno Fiaschetti e Paula Miraglia (Momentum) forneceram a análise contextual do Brasil. Jessica White e Irene Jay Liu (IFPIM) ofereceram apoio editorial e de pesquisa geral.

Sumário

The Protocol Gap: Brasil

Principais descobertas

- Uma esmagadora maioria dos sites de notícias no Brasil não possui qualquer política de proteção contra a raspagem por IA ou adota uma abordagem muito permissiva. Apenas 7,2% dos sites brasileiros de notícias desautorizam ao menos um rastreador de IA por meio de um arquivo robots.txt, embora a maior parte dos sites de notícias (75%) possua um arquivo robots.txt em seu site.

- Os sites no Brasil que implementam diretrizes em seus arquivos robots.txt concentram seus esforços no que parecem ser os crawlers de IA mais conhecidos, como os da OpenAI, Common Crawl, Google, ByteDance, Amazon, Apple, Meta e Huawei.

- Esses dados indicam que muitos sites não utilizam o robots.txt como mecanismo para sinalizar suas preferências quanto à raspagem de conteúdo por empresas de IA.

- Manter um arquivo robots.txt atualizado não é uma solução infalível, mas sim uma ferramenta prontamente disponível para que publishers declarem suas preferências quanto ao uso de seu conteúdo, alinhada à estratégia editorial e ao modelo de negócio dos veículos. Qualquer que seja sua posição, as organizações precisam de uma diretriz pública clara sobre como seu conteúdo pode ser utilizado para treinar novos modelos de IA.

Introdução

Nos últimos anos, veículos jornalísticos têm recebido um fluxo de tráfego de um novo tipo de visitante: os crawlers de IA. Do Googlebot ao "GPTbot" da OpenAI, ao "Claudebot" da Anthropic e ao "ExternalAgent" da Meta. Todos esses crawlers têm algo em comum: uma necessidade persistente de dados para treinar modelos de linguagem de grande porte (LLMs), desenvolver assistentes de IA e alimentar mecanismos de busca impulsionados por esta tecnologia. Esses novos visitantes são particularmente relevantes, considerando que cerca de 30% do tráfego global da web hoje vem de bots. Veículos jornalísticos são destinos atraentes, pois produzem informações oportunas e confiáveis que podem ajudar a melhorar a qualidade dos modelos e dos resultados de IA.

Essa crescente demanda por dados de IA apresenta novas oportunidades — mas também muitos desafios — para as organizações de mídia. Isso porque levanta questões jurídicas e éticas importantes relacionadas ao uso não autorizado de conteúdo, à violação de direitos autorais e a problemas de privacidade. Nesse arranjo, a própria sobrevivência do jornalismo está em jogo: sem mecanismos para sinalizar e aplicar permissões a crawlers de IA, organizações de mídia não conseguem proteger e monetizar de forma eficaz o valor do seu trabalho, que é raspado, reempacotado e redistribuído por sistemas de IA. Essa lacuna pode acelerar a crise de sustentabilidade enfrentada por veículos que já lidam com queda de visibilidade e tráfego vindos de plataformas digitais — o que muitos editores têm chamado de "apocalipse do tráfego" — e o colapso de modelos tradicionais de negócio baseados em publicidade.

Robots.txt é um arquivo de texto adicionado ao domínio de um site que contém instruções para mecanismos de busca e crawlers sobre quais páginas ou arquivos eles têm permissão para acessar. Apesar de suas limitações, o robots.txt é uma das poucas ferramentas gratuitas disponíveis para criadores de conteúdo sinalizarem se desejam que seus dados sejam usados por empresas de IA. O robots.txt já foi utilizado como prova em ações judiciais contra a raspagem e o rastreamento não autorizados no Canadá, nos Estados Unidos e no Reino Unido. O arquivo também pode servir como barganha em negociações de acordos de licenciamento, especialmente em contextos em que modelos de IA dependem fortemente de grandes volumes de dados coletados sem consentimento e sem compensação financeira.

Ainda que os veículos jornalísticos estejam cada vez mais conscientes desse desafio, são poucas as maneiras de gerenciar crawlers de IA — e nenhuma é totalmente infalível. Alguns sites estão adotando paywalls rígidos, enquanto outros iniciam longos processos judiciais alegando violação de direitos autorais. Respostas técnicas também estão surgindo: em julho de 2025, a provedora de infraestrutura e segurança de rede Cloudflare anunciou que suas proteções passariam a bloquear crawlers de IA

por padrão, reforçando um modelo baseado em permissões. Marketplaces de conteúdo — como o piloto lançado pela Microsoft, ou startups como ProRata.ai e Tollbit — e novas soluções de dados desenvolvidas por empresas como Flexolmo e OpenMined também estão emergindo como formas de dar aos produtores de conteúdo mais controle sobre como seu material é acessado, utilizado e remunerado por empresas de IA. Mas muitos ainda recorrem ao gerenciamento de raspagem de conteúdo por crawlers de IA por meio da implementação do Robots Exclusion Protocol (conhecido como robots.txt).

Apesar da crescente conscientização sobre a raspagem e o rastreamento não autorizados para treinamento de modelos de IA, as respostas das empresas de mídia para proteger seu conteúdo online continuam, na melhor das hipóteses, insuficientes. Veículos de pequeno e médio porte estão em desvantagem ainda maior porque não necessariamente possuem recursos para investir em infraestrutura sofisticada para detectar e bloquear crawlers de IA, nem têm poder de barganha para buscar acordos individuais de licenciamento com grandes empresas de IA — o Reddit, por exemplo, concordou em licenciar seu conteúdo ao Google para treinamento de IA em fevereiro de 2024.[1]

Sobre o projeto

Este relatório é o resultado de uma colaboração entre Journalism Relay Project, Momentum - Journalism and Tech Task Force, e International Fund for Public Interest Media (IFPIM). Neste primeiro relatório nacional, com foco no Brasil, buscamos mapear quantos sites jornalísticos estão utilizando o robots.txt para gerenciar crawlers de IA. O documento faz parte de um projeto de pesquisa mais amplo que envolve parceiros no Brasil, na Indonésia e na África do Sul, cujo objetivo é fortalecer a compreensão global sobre normas e práticas em evolução relacionadas ao acesso de sistemas de IA ao conteúdo jornalístico, com foco particular nos mercados-chave do Sul Global.

Por meio de pesquisa técnica e qualitativa, buscamos aumentar a conscientização entre organizações de mídia sobre como os sistemas de IA acessam seu conteúdo e ajudar a informar estratégias e políticas para gerenciar crawlers de IA de maneiras que melhor atendam aos interesses das redações. Desenvolvemos uma metodologia detalhada (ver anexo a este relatório) para aqueles que desejem replicar esta pesquisa sobre robots.txt em outros países e regiões.

```
# robots.txt for http://www.wikipedia.org/ and friends
#
# Please note: There are a lot of pages on this site, and there are
# some misbehaved spiders out there that go _way_ too fast. If you're
# irresponsible, your access to the site may be blocked.
#

# Observed spamming large amounts of https://en.wikipedia.org/?curid=NNNNNN
# and ignoring 429 ratelimit responses, claims to respect robots:
# http://mj12bot.com/
User-agent: MJ12bot
Disallow: /

# advertising-related bots:
User-agent: Mediapartners-Google*
Disallow: /

# Wikipedia work bots:
User-agent: IsraBot
Disallow:

User-agent: Orthogaffe
Disallow:

# Crawlers that are kind enough to obey, but which we'd rather not have
# unless they're feeding search engines.
User-agent: UbiCrawler
Disallow: /

User-agent: DOC
Disallow: /

User-agent: Zao
Disallow: /

# Some bots are known to be trouble, particularly those designed to copy
# entire sites. Please obey robots.txt.
User-agent: sitecheck.internetseer.com
Disallow: /

User-agent: Zealbot
Disallow: /

User-agent: MSIECrawler
Disallow: /

User-agent: SiteSnagger
Disallow: /

User-agent: WebStripper
Disallow: /

User-agent: WebCopier
Disallow: /

User-agent: Fetch
Disallow: /
```

Captura de tela do arquivo robots.txt usado pela Wikipedia.org

Embora o robots.txt permaneça o mecanismo mais amplamente disponível, esta pesquisa começa a revelar lacunas críticas no uso do arquivo por redações para gerenciar crawlers de IA, com foco em mercados do Sul Global. Neste primeiro relatório, mapeando mais de 4.000 sites de veículos jornalísticos no Brasil, constatamos que a grande maioria não fornece qualquer diretriz relacionada a crawlers de IA, apesar de muitas organizações de mídia já possuírem um arquivo robots.txt em seus sites.

À medida que empresas de IA continuam disputando a dominância do mercado e coletando grandes volumes de dados para alimentar seus sistemas, a necessidade de respostas mais informadas e coordenadas torna-se especialmente crucial para organizações de mídia no Sul Global. A adoção coordenada de robots.txt entre veículos jornalísticos pode ajudar a estabelecer padrões compartilhados que beneficiem toda a comunidade jornalística — de grandes redações a publishers independentes — ao definir diretrizes claras para empresas de IA. Esses simples documentos de texto continuam sendo uma ferramenta relevante para fortalecer organizações de mídia na era da IA, transformando o que poderia ser uma extração passiva de dados em uma decisão editorial ativa.

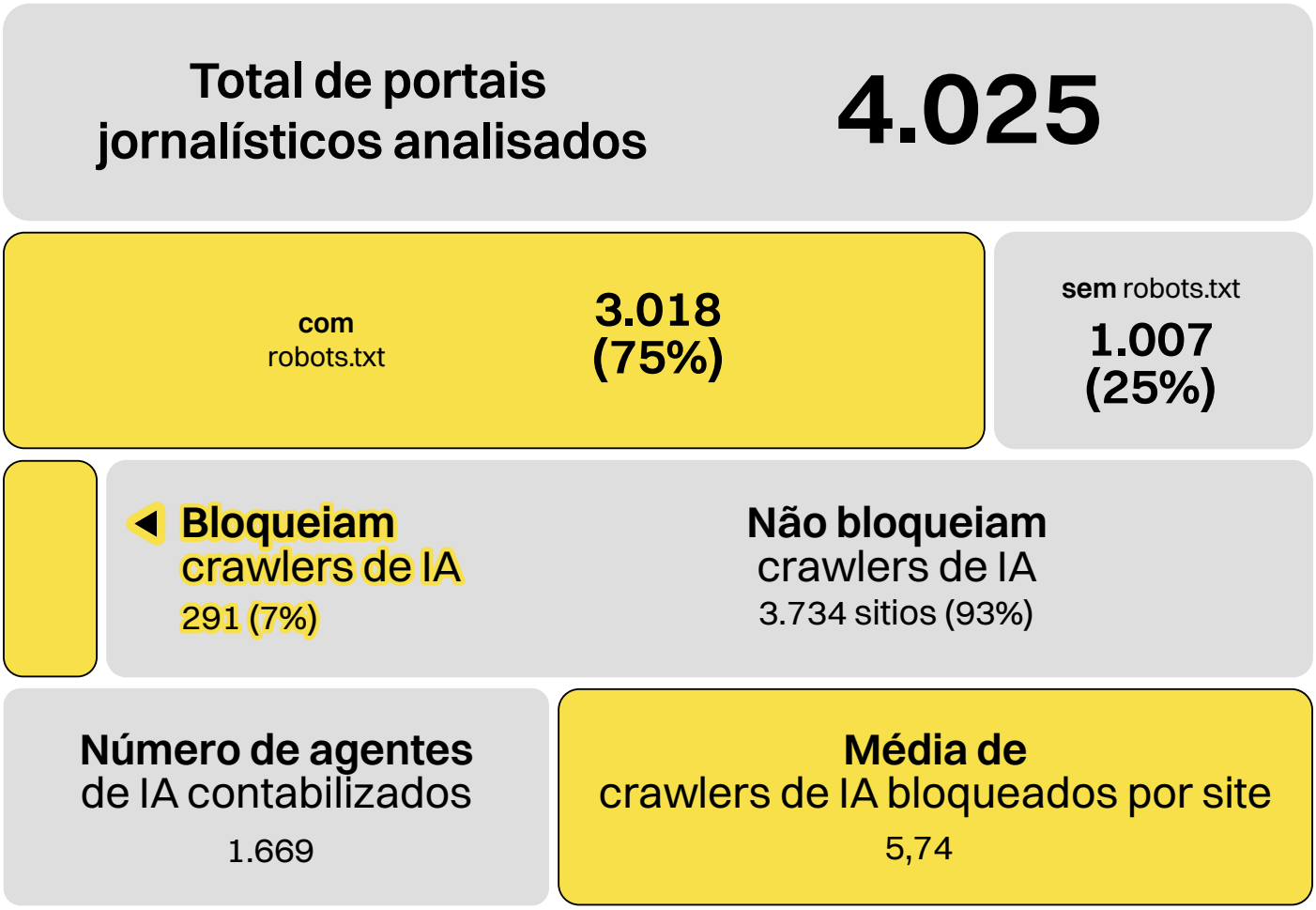
[1]Os dados sobre sites brasileiros foram baseados na pesquisa nacional mais recente do Atlas da Notícia (atlas.jor.br), de junho de 2025

Robots.txt: uma ferramenta subutilizada no Brasil

Nossa análise constatou que uma esmagadora maioria dos portais jornalísticos no Brasil não possui qualquer política de proteção à raspagem por IA ou adota uma abordagem muito permissiva, com diversos sites exibindo – quando o fazem – apenas os arquivos robots.txt básicos provenientes de seus sistemas de gerenciamento de conteúdo.

Apenas cerca de **7,2%** dos portais brasileiros tinham ao menos uma diretiva relacionada à IA, embora a maioria (**75%**) tenha sido identificada como possuindo um arquivo robots.txt anexado ao seu site.

Robots.txt entre publishers brasileiros: o que revelam os dados



Fonte: The Protocol Gap

Este é um forte indício de que a grande maioria dos sites desconhece as possibilidades de declarar publicamente seu nível de permissividade quanto à raspagem de conteúdo por IA, recorrendo, em vez disso, ao arquivo robots.txt padrão que vem por padrão em seus sistemas de gerenciamento de conteúdo.

Os portais brasileiros que implementam o bloqueio de IA adotam uma abordagem muito mais direcionada, concentrando seus esforços no que parecem ser os crawlers de IA mais conhecidos, em vez de adotar uma abordagem mais ampla. Os sites brasileiros bloqueiam, em média, apenas quatro agentes de IA (entre os que realizam o bloqueio).

Essa estratégia seletiva de bloqueio pode refletir diversos fatores: pode haver menor consciência sobre todo o espectro de crawlers de IA que operam online; os sites podem estar preocupados principalmente em bloquear apenas os sistemas de IA mais conhecidos; ou podem existir diferenças culturais e regulatórias que tornam o bloqueio abrangente de IA uma prioridade menor, em comparação com a abordagem mais cautelosa adotada por organizações de mídia em países como os Estados Unidos. Estudos adicionais são necessários para compreender quais fatores prevalecem.

Agentes bloqueados

Entre os sites que bloqueiam ao menos um agente de IA, os 10 bots mais bloqueados são:

Agente de IA	Descrição	% de bloqueio*
gptbot	Desenvolvido pela OpenAI. Os dados são usados para treinar modelos atuais e futuros, removendo conteúdo atrás de paywall, dados pessoais (PII) e dados que violem as políticas da empresa.	10,2
ccbot	Desenvolvido pela Common Crawl Foundation. Arquivo da web remontando a 2008. Citado em milhares de artigos de pesquisa por ano.	9,77
google-extended	Desenvolvido pelo Google. Usado para treinar os modelos generativos Gemini e Vertex AI. Não impacta a inclusão ou o ranqueamento de um site na Busca do Google.	9,53
claudebot	Desenvolvido pela Anthropic. Raspa dados para treinar LLMs e produtos de IA oferecidos pela Anthropic.	9,17
bytespider	Desenvolvido pela ByteDance. Faz download de dados para treinar LLMs, incluindo concorrentes do ChatGPT.	8,81
amazonbot	Desenvolvido pela Amazon. Inclui referências a sites rastreados ao exibir respostas via Alexa; não detalha claramente outros usos.	8,51
applebot-extended	Desenvolvido pela Apple. A Apple possui um agente secundário, Applebot-Extended, que é usado para treinar os modelos base que alimentam recursos de IA generativa em produtos da Apple, incluindo Apple Intelligence, Services e Developer Tools.	8,39
meta-externalagent	Desenvolvido pela Meta. O Meta-ExternalAgent rastreia a web para casos de uso, como treinar modelos de IA ou melhorar produtos, indexando conteúdo diretamente.	8,21
petalbot	Desenvolvido pela Huawei. Operado pela Huawei para fornecer serviços de busca e assistentes de IA.	3,82
universal_block	Usa a diretiva para bloquear todos os agentes externos. User-agent: * Disallow: /	1,92

Para uma lista de autoria e descrições dos bots, consulte: https://github.com/ai-robots-txt/ai.robots.txt/blob/main/table-of-bot-metrics.md

* Porcentagem calculada com base em todos os bloqueios observados (portanto, o cálculo não se refere à porcentagem de sites de notícias que bloqueiam).

Implicações para publishers no cenário brasileiro de desenvolvimento de IA

O uso limitado de robots.txt para impedir o crawling de IA evidencia uma lacuna na forma como os veículos jornalísticos brasileiros lidam com os impactos de tecnologias emergentes. Por um lado, as redações no Brasil têm adotado cada vez mais ferramentas de IA para diversas finalidades — desde traduzir artigos até produzir vídeos. De outro lado, o avanço da IA também está agravando desafios de sustentabilidade. Em um mercado de mídia historicamente concentrado em alguns grandes conglomerados, a falta de respostas proativas e coordenadas tende a intensificar desafios relacionados a relações assimétricas com plataformas digitais e dificuldades em negociar coletivamente com grandes empresas de tecnologia.

Muitos editores brasileiros têm expressado preocupação com o uso não autorizado de seu conteúdo, riscos éticos e queda no tráfego de seus sites — e manifestam o desejo de uma regulação justa e de celebrar contratos de licenciamento — mas estratégias coletivas e acordos concretos ainda não surgiram. Até o momento, nenhum acordo de licenciamento entre organizações de mídia e plataformas de tecnologia foi tornado público no Brasil. Em agosto de 2025, a Folha de São Paulo, um dos maiores jornais do país, entrou com uma ação judicial acusando a OpenAI de concorrência desleal e de violação de direitos autorais, alegando que a empresa utilizou seu conteúdo sem permissão ou pagamento para treinar seus modelos de IA. O recurso AI Overview também tem agravado as pressões financeiras, com evidências de impacto no tráfego de sites de notícias e em potenciais fontes de receita.

A falta de proteção ao conteúdo pode representar uma oportunidade perdida para garantir recursos e fortalecer a legitimidade do jornalismo. Do ponto de vista da sustentabilidade, a remuneração pelo uso de conteúdo jornalístico poderia ser uma fonte vital de financiamento, capaz de ao menos compensar parcialmente a perda de audiência e de receitas publicitárias resultantes da mudança digital nos padrões de consumo de informação. Diante da pressão financeira contínua, o ecossistema de mídia poderia encontrar nesta nova fonte de recursos uma forma de reforçar seu papel democrático.

Do ponto de vista da legitimidade, a extração de conteúdo jornalístico sem o devido reconhecimento de seu valor social pode agravar a erosão da posição do jornalismo como principal fonte de informação. No Brasil, estudos recentes revelam que — especialmente entre a população mais jovem — as plataformas de redes sociais se tornaram as principais ferramentas de acesso à informação, incluindo notícias e conteúdos políticos.

Agrava ainda mais esse cenário o fato de que a maior parte dessas empresas e tecnologias é sediada nos Estados Unidos e na China — o que significa que, tanto em sua concepção quanto em sua operação, essas ferramentas não só deixam de atender às necessidades e especificidades do Sul Global, como frequentemente reforçam e aprofundam assimetrias de poder entre Norte e Sul.

Apesar dos muitos desafios, o Brasil vive uma janela histórica para a discussão institucional dessas questões. Os três poderes — Executivo, Legislativo e Judiciário — têm avançado em medidas que podem redefinir o futuro do jornalismo.

O governo federal, ecoando aspirações mais amplas de outros países do Sul Global, tem defendido maior influência nos debates globais sobre soberania digital e regulação de IA. Exemplos recentes incluem a declaração do Grupo de Trabalho de Economia Digital do G20 sobre integridade da informação em 2024, bem como a mais recente declaração dos BRICS apresentada na cúpula realizada no Rio de Janeiro, em julho de 2025, que colocou a IA no topo da agenda do bloco. O engajamento do Executivo pode acelerar o avanço regulatório e abrir espaço para que a sociedade civil e o setor jornalístico influenciem essas discussões.

Enquanto isso, o Legislativo está avançando em projetos de lei importantes voltados à regulação das plataformas e sua relação com o jornalismo:

- PL 2630/2020 busca equilibrar a liberdade de expressão e o combate à desinformação, exigindo maior transparência algorítmica e determinando que plataformas que se beneficiam de conteúdo jornalístico remunerem aqueles que o produzem.

- PL de IA (PL 2338/2023) estabelece diretrizes gerais para o desenvolvimento e o uso de IA no Brasil. Seu objetivo é garantir práticas responsáveis que protejam direitos e sistemas democráticos. No que diz respeito especificamente à atividade jornalística, o texto afirma que o uso automatizado de obras — como extração, reprodução, armazenamento e transformação, incluindo a mineração de dados e de texto em sistemas de IA — por instituições de pesquisa, veículos de mídia, museus, arquivos e bibliotecas não constitui violação de direitos autorais. Essa exceção depende de as ferramentas serem usadas apenas quando necessário, sem prejudicar os interesses econômicos dos detentores de direitos, sem competir com a exploração normal das obras e sem serem usadas para fins comerciais sem compensação justa. A discussão sobre a obrigação de remunerar pelo uso de conteúdo protegido por direitos autorais é, porém, o ponto mais sensível do projeto. Grandes empresas de tecnologia têm pressionado deputados para evitar a obrigação de pagamento, tornando os parlamentares cautelosos em avançar com regulações mais rígidas devido às possíveis implicações econômicas e geopolíticas dessas medidas.

O Supremo Tribunal Federal também tem influenciado esse debate por meio de ações como o "Inquérito das Fake News" e a recente decisão que redefiniu a responsabilidade das plataformas por conteúdo gerado por usuários. Embora a posição da Corte tenha gerado preocupações sobre o risco de censura excessiva e o impacto na liberdade de expressão, ela tem influenciado o debate nacional sobre governança e regulação da IA.

Em meio a essas mudanças políticas, o setor de notícias permanece majoritariamente reativo aos novos paradigmas tecnológicos. Este estudo pode servir como um alerta. Com apenas 7,2% dos veículos brasileiros atualmente usando robots.txt para bloquear crawlers de IA, os achados destacam a urgência de aprofundar as discussões sobre apropriação de conteúdo e de desenvolver caminhos viáveis para o reconhecimento e a remuneração. Fortalecer a coordenação do setor em torno desses temas pode ser o primeiro passo rumo a um futuro digital mais sustentável para o jornalismo brasileiro.

Quando robots.txt falham, as redações pagam

Como o Google não diferencia entre seu mecanismo de busca tradicional e o recurso AI Overview, os publishers não conseguem usar o robots.txt para impedir que seu conteúdo seja utilizado em resumos gerados por IA sem, ao mesmo tempo, desaparecer completamente dos resultados de busca. Essa troca acarreta consequências econômicas significativas. Muitos veículos — inclusive no Brasil — relataram quedas acentuadas no tráfego de seus sites desde o lançamento do AI Overview, o que afeta diretamente sua capacidade de gerar receita. Isso ocorre porque os usuários têm menos probabilidade de clicar nos sites de notícias ao acessar as informações diretamente em um resumo gerado por IA. Em setembro de 2025, a Cloudflare anunciou a Content Signals Policy, projetada para se integrar ao arquivo robots.txt e oferecer aos proprietários de sites maior controle para bloquear crawlers usados para treinamento de IA, ao mesmo tempo em que permanecem visíveis na busca tradicional. No entanto, ainda não está claro se essa medida será suficiente para conter a perda de tráfego nos sites.

Conclusão e próximos passos

Este estudo serve como ponto de partida para reflexões mais profundas sobre a complexa relação entre o jornalismo e as plataformas digitais no Sul Global — especialmente no que diz respeito ao uso não autorizado de conteúdo jornalístico por ferramentas de IA generativa e ao desenvolvimento de estratégias para compensar de forma justa os veículos de mídia.

Apesar de suas limitações — em especial as restrições técnicas do robots.txt como ferramenta de proteção de conteúdo e a dependência exclusiva de dados públicos — os achados apontam para um sentimento mais amplo de despreparo, ou até mesmo de impotência, entre os publishers brasileiros diante da captura de seu conteúdo por empresas de IA. A conscientização sobre como poucos veículos brasileiros protegem ativamente seu conteúdo por meio de robots.txt também pode marcar um ponto de inflexão, afastando o setor da fragmentação e da falta de engajamento e aproximando-o de uma discussão coletiva sobre a regulação da IA, estimulando uma nova onda de reflexão estratégica na indústria jornalística.

Este estudo representa o primeiro capítulo de uma parceria entre Journalism Relay Project, Momentum – Journalism and Tech Task Force e International Fund for Public Interest Media (IFPIM), destinada a construir um campo crítico de conhecimento e contribuir para discussões sobre caminhos para a sustentabilidade do jornalismo a partir da perspectiva do Sul Global.

Os próximos passos incluem:

1. **Pesquisas futuras:** Com base nos dados e conclusões apresentados aqui, planejamos avançar com uma análise qualitativa do cenário brasileiro por meio de entrevistas com publishers locais. Nosso objetivo é reunir novos insights que nos ajudem a compreender com mais clareza e profundidade os dilemas enfrentados pelo jornalismo no Brasil em relação à captura de conteúdo e à busca por estruturas regulatórias justas para o uso de seu material.

2. **Ferramentas para publishers:** A decisão de bloquear ou não crawlers é individual e deve estar alinhada à estratégia de cada veículo.

3. **Parceria transnacional:** A parceria também pretende replicar a metodologia utilizada neste estudo para examinar os ecossistemas midiáticos da Indonésia e da África do Sul — outros dois países do Sul Global que, assim como o Brasil, compartilham características semelhantes: mercados de mídia historicamente concentrados, relações desiguais com plataformas digitais e dificuldades para negociar coletivamente com empresas de Big Tech.

Incentivamos todos os pesquisadores que desejem replicar este estudo em outros mercados e sugerimos que consultem a metodologia completa apresentada a seguir.

Metodologia

1.1.Origem dos dados

Este projeto utilizou verificação programática de arquivos para identificar a existência de parâmetros de bloqueio para agentes de IA nos arquivos robots.txt de 4.023 organizações jornalísticas do Brasil. Os dados sobre websites no Brasil foram baseados na pesquisa nacional mais recente do Atlas da Notícia (atlas.jor.br), de junho de 2025. As principais variáveis consideradas foram os nomes das organizações jornalísticas, seus endereços na web e suas localizações.

A metodologia concentrou-se em examinar esses arquivos em busca de padrões específicos e de sua relação com crawlers de IA, com base no parâmetro padrão **Disallow:/** em arquivos robots.txt. Ao consultar uma base de dados open source de identificadores conhecidos de crawlers de IA — incluindo os de grandes empresas de tecnologia como Google, Meta, Amazon, Apple, Anthropic e OpenAI — a análise identificou quando websites tentavam sinalizar restrições específicas a crawlers de IA quanto ao uso de seu conteúdo.

Esse aspecto da metodologia foi particularmente relevante diante das preocupações recorrentes sobre sistemas de IA extraindo conteúdo da web sem permissão explícita. Vale destacar que arquivos robots.txt são instrumentos não vinculantes, que não impedem tecnicamente a coleta de dados por bots, mas apenas indicam a política de um site em relação à agregação automatizada de conteúdo.

Ao examinar um website, o sistema primeiro garantia a formatação adequada da URL antes de tentar acessar o arquivo robots.txt. Em seguida, aplicava protocolos de tratamento de erros e de tempo limite (timeout) para lidar com diversos cenários reais, como tempo de resposta do servidor, arquivos ausentes, problemas de conexão ou bloqueios de DNS, criando assim um sistema abrangente de registro (logging).

Para considerar servidores de resposta lenta que poderiam precisar de tempo para "acordar" ("wake up"), o sistema implementou um mecanismo de nova tentativa com um atraso de 3 segundos quando arquivos robots.txt não eram encontrados na primeira tentativa. Essa abordagem aumentou significativamente a confiabilidade da detecção, mantendo tempos de processamento eficientes para servidores responsivos.

A lógica de análise distinguiu entre diretivas restritivas (**Disallow: /** ou **Disallow: ***) e diretivas permissivas (Allow: ou declarações Disallow: vazias), garantindo uma classificação precisa das políticas de bloqueio de IA.

Os sites foram categorizados em um espectro que ia do permissivo ao super restritivo, com base no número e no tipo de agentes de IA bloqueados.

Websites com arquivos robots.txt malformados ou incompletos foram, em geral, classificados como permissivos para crawlers de IA. Isso incluiu sites com diretivas **Disallow:** vazias (que tecnicamente permitem acesso) e aqueles que utilizavam regras de bloqueio apenas parciais, restringindo URLs ou diretórios específicos, mas deixando a maior parte do conteúdo sem restrições. Na ausência de diretivas de bloqueio abrangentes, esses sites foram efetivamente tratados como permitindo acesso de agentes de IA ao seu conteúdo.

Realizamos uma checagem manual de 90 resultados aleatórios e constatamos que a maior parte dos achados estava, de fato, correta. Apesar dessas salvaguardas metodológicas robustas, por excesso de cautela estimamos que persiste uma taxa de erro de 3% devido a fatores técnicos, como problemas de resolução de DNS, timeouts de servidores e falhas temporárias de conectividade, que podem resultar em falsos positivos ou, em alguns casos, falsos negativos na detecção de robots.txt.

1.2 Coleta de dados

Uma vez que o conteúdo do arquivo robots.txt havia sido recuperado com sucesso, a análise avançava para o reconhecimento de padrões. O sistema processava o conteúdo linha por linha, identificando diretivas User-agent e comparando-as à base de dados de crawlers de IA conhecidos (buscando o parâmetro /Disallow).

Esse processo de correspondência de padrões foi implementado com comparação case-insensitive para capturar variações na forma como os nomes dos crawlers poderiam estar escritos, tornando a análise menos suscetível a redundâncias. Os resultados dessa análise foram estruturados para fornecer insights sobre a postura de cada website em relação aos crawlers de IA.

Para cada website analisado, o sistema determinou se existia um arquivo robots.txt, se ele continha diretivas de bloqueio e, especificamente, quais crawlers de IA conhecidos estavam sendo bloqueados no parâmetro **/Disallow**.

Essa abordagem estruturada de coleta e análise permitiu realizar de forma rápida (média de 90 minutos de execução do código) exames em larga escala sobre como diferentes websites tratavam o acesso de crawlers de IA.

A coleta e a análise foram realizadas em R.

[2]Atlas da Notícia, site - https://atlas.jor.br
[3]Lista de crawlers de IA conhecidos, 2025, repositorio de Github – ai.robots.txt

[4]OECD (2025), "Intellectual property issues in artificial intelligence trained on scraped data", OECD Artificial Intelligence Papers, No. 33, OECD Publishing, Paris, https://doi.org/10.1787/d5241a23-en.

1.3 Classificação dos dados

À medida que os dados eram coletados a partir dos arquivos robots.txt, o algoritmo automaticamente realizava classificações com base no nível de permissividade discretamente atribuído a cada site.

Em vez de utilizar limites absolutos arbitrários, o sistema primeiro calcula a média basal do número de agentes de IA bloqueados pelos sites que efetivamente implementam políticas de bloqueio. Essa média serve como referência para determinar o que constitui um comportamento de bloqueio normal, moderado ou excessivo no conjunto específico de dados analisados.

O sistema de classificação, então, cria cinco faixas distintas usando múltiplos dessa média calculada.

Essa metodologia garante que a classificação permaneça significativa independentemente dos números absolutos em qualquer conjunto de dados, pois ela se adapta à distribuição real dos comportamentos de bloqueio observados. Por exemplo, se um site classificado como médio bloquear três agentes de IA, um site que bloqueie trinta agentes seria classificado como super-restritivo; porém, se a média do conjunto de dados fosse de dez agentes, esse mesmo site que bloqueia trinta agentes seria classificado como restritivo.

Classificação	**Agentes de IA bloqueados**	**Descrição**
Política permissiva para agentes de IA	0 agentes de IA	Sites que não impõem restrições à crawlers de IA
Pouco restritivo	1x a 2x a média	Sites com comportamento de bloqueio típico ou ligeiramente acima da média
Restritivo	2x a 5x a média	Sites que bloqueiam de duas a cinco vezes a média
Muito restritivo	5x a 10x a média	Sites que bloqueiam de cinco a dez vezes a média
Super restritivo	10x+ a média	Sites com políticas de bloqueio excepcionalmente abrangentes

1.4 Agregação dos dados

Por fim, os dados foram agregados em um conjunto de indicadores que pudessem representar o conjunto analisado como um todo.

Métrica	**Valor / unidade**	**Dados observados**
total_observations	data / websites	data
block_ai	data / websites	data
not_block_ai	data / websites	data
pct_not_blocked	data / percent	data
pct_blocked	data / percent	data
yes_robots_txt	data / websites	data
no_robots_txt	data / websites	data
pct_robots_txt	data / percent	data
pct_no_robots_txt	data / percent	data
ai_agents_count	data / ai agents	data
avg_agents_per_blocking_site	data / ai agents	data

/members/api/comments/counts/ Disallow: /r/ Disallow: /webmentions/receive/ # Allow rules User-agent: Google Disallow rules # robots.txt generated by Robots.txt Helper AI2Bot Disallow: / User-agent: Ai2Bot-Dolma Disallow: / U aiHitBot Disallow: / User-agent: Amazonbot Disallow: / Us Andibot Disallow: / User-agent: anthropic-ai Disallow: / Us Applebot Disallow: / Use -agent: Applebot-Extended Disal User-agent: bedrockbot Disallow: / User-agent: Brightbot User-agent: Bytespider Disallow: / User-agent: CCBot Disa User-agent: ChatGPT-User Disallow: / User-agent: Claude Disallow: / User-agent: Claude-User Disallow: / User-age Disallow: / User-agent: THE PROTOCOL GAP: ClaudeBot D User-agent: cohere-ai Disallow: / User-agent GoogleOthe Disallow: / User-agent:cohere-training-data-crawler Disa User-agent: Cotoyogi Disallow: / User-agent: Crawlspace User-agent: BRAZIL Disallow: / User-agent: DuckAssistBo User-agent: EchoboxBot Disallow: / User-agent: Faceboo / User-agent: Factset_spyderbot Disallow: / User-agent: F Disallow: / User-agent: FriendlyCrawler Disallow: / User-a Google-CloudVertexBot Disallow: / User-agent: Google-E Disallow: / User-agent: GoogleOther Disallow: / User-agen GoogleOther- Março 2026 : / User-agent: GoogleOther-Vi / User-agent: GPTBot Disallow: / User-agent: iaskspider/2 User-agent: ICC-Crawler Disallow: / User-agent: Imagesif / User-agent: img2dataset Disallow: / User-agent: ISSCyb Disallow: / User-agent: Kangaroo Bot Disallow: / User-age meta-externalagent Disallow: / User-agent: Meta-Externa