members/api/comments/counts/ Disallow: /r/ Disallow:
webmentions/receive/ # Allow rules User-agent: Google
Disallow rules # robots.txt generated by Robots.txt Helpe
AI2Bot Disallow: / User-agent: Ai2Bot-Dolma Disallow: / U
iHitBot agent: Amazonbot Disallow: / Us
ndibot agent: anthropic-ai Disallow: / U
pplebo -agent: Applebot-Extended Disa
User-agent: bedrockbot Disallow: / User-agent: Brightbo
User-agent: Bytespider Disallow: / User-agent: CCBot D
User-agent: ChatGPT-User Disallow: / User-agent: Claud
Disallow: / User-agent: Claude-User Disallow: / User-age
Disallow: / User-agent ClaudeBot
User-agent: cohere-ai Disallow: / User-agent GoogleOth
Disallow: / User-agent:cohere-training-data-crawler Dis
User-agent: Cotoyogi Disallow: / User-agent: Crawlspac
User-agent: / User-agent: DuckAssistB
User-agent: EchoboxBot Disallow: / User-agent: Faceboo
Disallow: / User-agent: Factset_spyderbot Disallow: / Use
FirecrawlAgent Disallow: / User-agent: FriendlyCrawler I
User-agent: Google-CloudVertexBot Disallow: / User-age
Google-Extended Disallow: / User-agent: GoogleOther D
User-agent: GoogleOther- : / User-agent:
GoogleOther-Video Disallow: / User-agent: GPTBot Disal
User-agent: iaskspider/2.0 Disallow: / User-agent: ICC-C
Disallow: / User-agent: ImagesiftBot Disallow: / User-age
img2dataset Disallow: / User-agent: ISSCyberRiskCrawle
User-agent: Kangaroo Bot Disallow: / User-agent: meta-e

CTRL + J

THE PROTOCOL GAP:

SOUTH AFRICA

August 2026

Gordon Institute of Business Science
University of Pretoria

International Fund for Public Interest Media

Journalism Relay Project

This report is a joint publication of the Journalism Relay Project, the Media Leadership Think Tank (MLTT) at the Gordon Institute of Business Science (GIBS) in Johannesburg, and the International Fund for Public Interest Media (IFPIM).

Sérgio Spagnuolo (Journalism Relay Project) developed the methodology and led the data analysis.

Ompha Tshamano and Michael Markovitz (MLTT) provided contextual analysis for South Africa.

Jessica White and Irene Jay Liu (IFPIM) provided overall editorial and research support.

# Contents

# The Protocol Gap: South Africa

## Key Findings

- As of April 2026, 30.4% of South African news websites disallow at least one AI crawler through their robots.txt file, even though most (74.1%) have a robots.txt file on their website. South African publishers that implement robots.txt directives to manage AI bots have risen modestly from 29.7% recorded in December 2025.
- Sites that implement robots.txt directives target an average of 6 AI bots, indicating that publishers that take action tend to adopt broad blocking strategies rather than singling out individual crawlers. Frequently restricted bots include several well-known crawlers from Common Crawl Foundation, ByteDance, Google, OpenAI, Amazon, and Anthropic.
- While South Africa's 30.4% is relatively high compared to Brazil and Indonesia, examined under the same methodology, adoption remains concentrated among large, well-resourced publishers. The uneven distribution of AI-scraping policies points to significant disparities in institutional resources, technical capacity, referral traffic dependency, and strategic positioning across a diverse range of South African publishers.
- Amid ongoing uncertainty about the legal and market rules governing AI-driven content extraction, robots.txt remains a simple, available way for newsrooms to state how they want their content used, in line with their content strategy and business model.

## Introduction

South Africa is engaged in an ongoing debate about how artificial intelligence should interact with journalistic content. For a news media industry under significant financial pressure, the rapid development of AI is adding a new sense of urgency.

In late 2025, the landmark Media and Digital Platforms Market Inquiry (MDPMI) led by South Africa's Competition Commission recognised the adverse effects of platform power on the sustainability of local media, and echoed growing concerns about AI companies extracting publisher content to train AI models and generate content, often without giving credit or consent to the original producers of information. For smaller outlets in particular, the loss of referral traffic due to AI-generated summaries is a direct threat to revenue and survival. The MDPMI concluded that referral traffic from AI chatbots to third-party websites, including news media, was 'insignificant.'

Amid ongoing global uncertainty about the legal and market frameworks governing AI-related content extraction, the decisions publishers make now about how AI crawlers access their content could have long-lasting consequences. Without mechanisms in place to signal and enforce permissions on AI crawlers, media organisations are unable to effectively protect and monetise the value of their work as it gets scraped, repackaged, and redistributed by AI systems. This gap risks accelerating the sustainability crisis facing media outlets that are already grappling with declining visibility and traffic from digital platforms, which many publishers have already called a 'traffic apocalypse.'

News sites are responding to this challenge in several ways. Some sites are implementing hard paywalls, while others are launching lengthy lawsuits claiming copyright infringement. Technical responses are also emerging: in July 2025, the internet infrastructure provider Cloudflare announced it would block AI crawlers by default under a permission-based model, and in July 2026 it introduced a default block on "multi-purpose" crawlers and more options for websites to manage AI crawlers by function (search,

agent, training). Content marketplaces like Microsoft's pilot or startups such as ProRata.ai and Tollbit, and new data solutions such as Flexolmo, and OpenMined and Mozilla Data Collective are also offering publishers more control over how their content gets accessed, used, and compensated by AI companies. But many publishers still rely on managing bots through the Robots Exclusion Protocol (known as robots.txt).

While robots.txt remains a widely available mechanism, this study uncovers significant gaps in newsroom adoption of robots.txt to manage AI crawlers, with a lens on Global South markets. Following a first report focused on Brazil, this second report is based on an analysis of 263 South African news websites. Our findings reveal uneven engagement with robots.txt as a mechanism to govern AI crawler access: while most South African publishers have implemented robots.txt in some form, only a third have adopted targeted AI directives.

Decisions about whether to allow or disallow AI crawlers are shaped not only by technical considerations but also by trade-offs between visibility and control, audience reach and content protection, and short-term sustainability and long-term values. Where restrictive practices are observed, they are applied inconsistently, with only modest change over time. These patterns suggest that publishers are actively navigating but not yet fully resolving the challenges posed by AI-driven content extraction.

**What is robots.txt?**

Robots.txt is a text file added to the root domain of a website that contains instructions to search engines and web crawlers about which pages or files they are allowed to access. This includes directives to AI bots – automated programmes that systematically browse and collect data from websites to train large language models and power AI assistants and search engines.

Despite its limitations, robots.txt is one of the few free tools available to content creators to signal whether they want their data to be used by AI companies. While not legally binding, robots.txt has already been used as evidence in legal complaints against unauthorised crawling and scraping in Canada, the United States, and the United Kingdom. The file can also serve as leverage in licensing negotiations, especially in contexts where AI models rely heavily on large volumes of unconsented, uncompensated data.

As such, robots.txt functions as a mechanism for managing crawler access but also as a strategic tool through which publishers can signal consent, refusal, or conditional use. However, its effectiveness depends on voluntary compliance by AI developers, and its role within broader governance frameworks remains contested.

South Africa offers a particularly instructive case towards understanding the opportunities and challenges for developing more informed responses. The country's media ecosystem is shaped by high ownership concentration, uneven digital capacity, and deep inequalities rooted in its political economy. These structural conditions mean that the capacity to respond to AI scraping is not evenly distributed. The MDPMI report confirmed that larger publishers with diversified revenue and strong direct audiences are better positioned to restrict crawler access, while community, vernacular, and smaller independent outlets face greater pressure to remain open in order to maintain visibility and traffic.

Our findings signal that this asymmetry plays out in the unequal adoption of robots.txt between larger and smaller publishers, yet more research is needed to better understand the practical barriers and challenges impacting publishers' ability to make informed decisions based on their content strategies and business models. The sections that follow present the study's empirical findings in detail and explore their implications for media sustainability, market structure, and AI governance in South Africa.

**About this project**

This report is the result of a collaboration between the Journalism Relay Project, the Media Leadership Think Tank (GIBS) and the International Fund for Public Interest Media. In this country report focusing on South Africa, we sought to map how many news sites use robots.txt to manage AI crawlers.

It is part of a broader research project that involves partners in Brazil, Indonesia, and South Africa and aims to strengthen our global understanding of evolving norms and practices around AI access to journalistic content, with a particular focus on key markets in the Global South.

Through technical and qualitative research, we aim to increase awareness among media organisations on how AI systems access their content and help to inform strategies and policies to manage AI crawlers in ways that best serve newsroom interests. We have developed a detailed methodology (see Methodology section in this report) for all those who may want to replicate this research on robots.txt in other countries and regions.

# Robots.txt Among South African Publishers: What the Data Tells Us

**Total news websites analysed** **263**

News websites **with** robots.txt: **195 websites, 74.1%**

**without** robots.txt **68 websites, 25.9%**

**Blocking** AI crawlers
80 websites, 30.4%

**Not blocking** AI crawlers
183 websites, 69.6%

**Number of** AI agents counted
**482**

**Average number** of AI crawlers blocked per site
**6**

Source: The Protocol Gap

## Blocked Agents

Among the websites that do block at least one AI agent, the top 10 most blocked bots are:

| AI agent | Description | % of blocks* |
|---|---|---|
| ccbot | Common Crawl Foundation - Provides open crawl dataset, used for many purposes, including Machine Learning/AI. | 12.7 |
| bytespider | ByteDance - Downloads data to train LLMS, including ChatGPT competitors. | 5.2 |
| google-extended | Google - LLM training. | 5.2 |
| gptbot | OpenAI - Scrapes data to train OpenAI's products. | 5.2 |
| amazonbot | Amazon - Service improvement and enabling answers for Alexa users. | 5.0 |
| claudebot | Anthropic - Scrapes data to train Anthropic's AI products. | 4.8 |
| anthropic-ai | Anthropic - Scrapes data to train Anthropic's AI products. | 2.9 |
| chatgpt-user | OpenAI - Takes action based on user prompts. | 2.9 |
| claude-web | Anthropic - Undocumented AI Agents. | 2.9 |
| diffbot | Diffbot - Aggregates structured web data for monitoring and AI model training. | 2.9 |

For a list of authorship and descriptions of bots, see: https://github.com/ai-robots-txt/ai.robots.txt/blob/main/table-of-bot-metrics.md
* Percentage calculated from all observed blocks (not the percentage of news websites blocking).

# Implications for Publishers in South Africa's Evolving AI Landscape

South Africa's media ecosystem is characterised by significant ownership concentration, uneven digital capabilities, and persistent inequalities rooted in the apartheid-era political economy.[1] Our findings reflect a sector still navigating an uncertain and fragmented strategic environment around AI.

Among the sampled South African news websites, 74.1% had robots.txt files, but only 30.4% had implemented any AI-specific crawler directives, a modest increase from 29.7% in December 2025. This relatively flat trajectory contrasts with preliminary findings from Indonesia, where AI blocking more than doubled from 5.8% to 13% over a similar period, and Brazil, where it rose from 5% to 7%. Even so, South Africa's blocking rate remains substantially higher than both in absolute terms.

This combination of a relatively high but slow-growing blocking rate reflects a sector still navigating an uncertain and fragmented strategic environment around AI governance. The gap between broad robots.txt adoption (74.1%) and targeted AI-specific directives (30.4%) could be a result of publishers using the default robots.txt that came with their website. It could also indicate that while many publishers engage with basic search engine governance, fewer are proactively addressing the challenges posed by AI scraping. Decisions about whether to allow, disallow, or selectively manage AI scraping are shaped not only by technical considerations, but also by uncertainty regarding legal rights, bargaining power, audience dependency, and future commercial positioning within emerging AI-driven information ecosystems.

One factor contributing to this uncertainty, the status of South Africa's Copyright Amendment Bill (CAB), was partially resolved on 26 June 2026, when the Constitutional Court upheld the Bill's new fair use provisions but struck down some of its educational copying exceptions as unconstitutional. Because of that second finding, the Bill has to be returned to Parliament for reconsideration before it can be signed into law, with no deadline set for this process.

This uncertainty extends beyond copyright alone and includes the absence of clear licensing frameworks, compensation mechanisms, and sector-wide governance arrangements for AI and journalism. In the absence of clear legal and market framework(s) governing AI-related content extraction, robots.txt has increasingly become a *de facto* governance mechanism for publishers. Although it carries no formal legal force and depends largely on voluntary compliance by AI developers, it enables publishers to signal consent, refusal, or conditional access within a highly asymmetrical digital environment.

**Growing Legal Scrutiny of AI Training Practices**

International legal developments have highlighted the growing scrutiny of AI training practices and the provenance of training data, including disputes over copyrighted works, unauthorised scraping, and fair use interpretations.

The Bartz v Anthropic case was a U.S. class action in which authors challenged Anthropic's use of pirated books (including the works of South African authors Nadine Gordimer and Zakes Mda) to train Claude. While their involvement in the matter does not create a legal precedent, it is nonetheless significant: it foregrounds the material exposure of South African rights-holders to global AI training practices and brings local visibility to the legal and protocol gaps under discussion, particularly in relation to data provenance, consent, and the extraction of value from copyrighted works in the absence of enforceable safeguards.

Investigations such as The Atlantic's AI Watchdog project further echo these concerns, revealing how similar opaque training practices sweep in books, music and other creative works without consent or compensation, reinforcing patterns of 'cultural free-riding' that South African policymakers can no longer treat as distant or hypothetical.

1 The Competition Commission's MDPMI final report found that 'despite its diversity, with 380 print publishers, 5 community TV stations and 207 community radio stations, mainstream media ownership in South Africa reflects a global trend of consolidation, with a small number of companies dominating across platforms.'

The uncertainty surrounding copyright law in South Africa has been compounded by the fragility and uncertainty of the country's broader AI policy environment. The Draft National Artificial Intelligence Policy Framework was withdrawn from public consultation, following concerns about fictitious AI-generated references contained in the document. Although the draft policy was welcomed by some for its developmental ambitions and global positioning, critics pointed to its insufficient engagement with value extraction, compensation and the treatment of copyrighted content in AI systems. Some also questioned whether the policy's proposed governance architecture was realistic, sufficiently adaptive to rapid technological change, or adequately grounded in South Africa's institutional and infrastructural constraints. These gaps may undermine the policy's credibility, but they do not negate its legitimacy as a necessary starting point for public consultations and policy developments. The draft policy deferred these issues to existing copyright frameworks and the then-pending ruling on the CAB, thereby leaving the position of journalism within emerging AI markets unresolved.

Without policy guidance, tech companies will continue to determine how AI is developed, implemented, monetised, and used. Furthermore, the lack of guidance will result in publishers being responsible for establishing transparent, ethical guidelines for and education about AI use, without support or guidance from regulators. The impact of generative AI on the sustainability of news media has now reached the multilateral agenda through UNESCO's ongoing Draft Guidance on Fair Compensation for News, reflecting growing recognition that threats to journalism's economic foundations carry wider risks for information integrity, democratic participation, and fundamental rights.

**A Strategic Dilemma: To Protect Content or Remain Visible?**

Since Google has not clearly separated AI Overviews from its traditional search engine bot (Googlebot), publishers have been facing a difficult trade-off between protecting their content and staying visible to audiences in the dominant search platform. While publishers can disallow Google-extended in robots.txt to block some AI training use of their content (e.g. for Gemini-based AI models), if they want to reliably block use of their content for AI Overviews, they would need to remove their pages from Google Search results too. This trade-off has already had economic consequences.

However, several emerging responses may offer publishers greater leverage and bargaining power:

- In June 2026, the UK Competition and Markets Authority (CMA) imposed new requirements on Google Search, giving news organisations more control over whether their content is used by AI features such as AI Overviews and AI mode. This decision sets a global precedent as Google said it is testing distinct opt-out controls in the UK "before rolling them out to website owners globally."
- Other proceedings are underway: in April, Brazil's competition authority (CADE) approved the launch of an inquiry into Google's conduct and its impact on journalism.

Other tools are also emerging, such as Cloudflare's Content Signals Policy and new options to manage AI bots, which could give site owners stronger controls to block crawlers for AI training while still allowing them to be visible in traditional search.

# An industry grappling with the impact of AI

According to audience-measurement data reviewed by the MLTT in July 2026, South Africa's leading news sites lost, on average, one-fifth of their total daily page views between May 2025 and May 2026. This mirrors the international picture where 42 of the world's top 50 English-language news sites declined over the same period, with a median fall of 17%.

The MDPMI report emphasised that AI chatbots and large language models have scraped news content without compensation, using it to train AI systems and generate responses to user queries. The Inquiry identified growing concerns around AI-related content extraction, bargaining asymmetries, and the potential disintermediation of publishers through AI-driven interfaces. Most importantly, the Inquiry recognised that publishers are increasingly forced to trade off protecting their content from uncompensated extraction against maintaining visibility, traffic, and revenue in AI-mediated and platform-dependent discovery environments. As access to news by 'audiences' becomes increasingly shaped by AI-generated summaries and search interfaces, publishers face structural constraints in resisting content use without risking exclusion from digital ecosystems.

The Commission found that AI-powered search summaries can substitute for visits to news websites and that tying their use to general search indexing effectively forces publishers to make their content available. It concluded that this 'would constitute unfair competition' and was 'unlikely to constitute fair use' because the use competes commercially with the copyright holder and affects the market value of the work.

The MDPMI report states that the central issue requiring remedial action is empowering South African media to opt out of AI training and content use. The Commission therefore imposed platform-specific remedial actions aimed at strengthening publisher controls and capacity:

- **Google:** Biannual AI training for non-mainstream publishers on controls and opt-outs for training and grounding, alongside an experimental African News Innovation Forum.
- **Microsoft:** Annual training on AI controls and opt-outs; if its Publisher Content Marketplace is commercialised, it must be extended to South African publishers on equivalent terms.
- **Meta:** Four annual publisher-support workshops, including AI digital literacy and training on Meta AI tools and the robots.txt opt-out.
- **OpenAI and X.AI:** Biannual training for all SA news publishers on controls and opt-outs relating to AI training and grounding.

The remedial framework itself recognises capacity constraints: the Commission identified periodic training and technical support as an effective and practical remedy, 'particularly for smaller media that lack the resources.'

While the MDPMI remedies acknowledge growing concerns about AI-related content extraction and platform power, they do not fully account for how uneven referral traffic dependency shapes publishers' practical capacity to exercise content control.

This challenge is compounded by the limited volume of traffic currently being returned to publishers through AI-mediated discovery systems. Annexure 5 of the MDPMI report recorded that popular AI chatbots referred just under 2 million visits to 222 South African news websites between January and May 2025. However, these referrals accounted for less than 1% of the 256 million chatbot visits recorded over the same period. The findings suggest that, despite growing concerns about AI-related content extraction, publishers are currently receiving only a small share of the value generated through chatbot-mediated information access. This reinforces concerns about disintermediation and the weakening of traditional business models that have historically relied on audience traffic and advertising revenue.

As a result, the costs and benefits of restricting AI access are not experienced equally across the sector. In line with international trends, larger news organisations in South Africa tend to have stronger direct audiences, greater technical capacity, and more diversified revenue streams. This has enabled them to move early to either block AI agents or negotiate licensing agreements. An earlier analysis published by The Outlier in August 2025 found that larger, more established publishers with successful print-to-digital transitions had blocked most AI agents, namely GPT, Claude, Google AI and Perplexity. By contrast, digital natives were not blocking AI agents, whilst the national broadcaster (SABC) and other broadcasters had no robots.txt files. Digital-native and community outlets, many of which are more reliant on platform-mediated discovery, may face greater pressure to remain open to AI scraping to sustain visibility, audience reach, and revenue. This suggests that the practical ability to restrict AI access may be unevenly distributed across the media ecosystem. For these outlets, openness to AI scraping increasingly appears to function as a direct or indirect condition for visibility, rather than a 'genuinely' voluntary choice.

This correlates with the April 2026 robots.txt dataset, which indicates that the divergence is not only sectoral but deeply structural:

- Larger media groups display the most restrictive AI policies by volume and breadth. These publishers combine strong brand recognition, subscription buffers, and in-house technical capacity, allowing them to limit AI scraping while maintaining audience reach through direct traffic and audience loyalty.
- Local media and regional outlets, many of which are often owned by large media groups, and regional publishers often operate on thin financial margins. A large cluster of hyperlocal titles blocks a small selection of blocks, signalling partial awareness without the capacity or leverage to implement comprehensive protections.
- By contrast, several digital-native publishers, despite significant national visibility and strong technical capacity, remain comparatively permissive.
- At the other end of the spectrum, many community and civic outlets block no agents at all, effectively remaining fully exposed.

A publisher's decision to remain more or less permissive towards AI crawlers may reflect a range of factors, including audience growth strategies, referral-traffic dependency, public interest considerations, or differing assessments of the risks and opportunities associated with AI-mediated discoverability. For some publishers, openness may function less as straightforward consent than as a strategic trade-off shaped by visibility, growth and sustainability considerations.

However, the motivations underlying these decisions require further qualitative research and should not be inferred solely from robots.txt behaviour.

# Conclusion and Next Steps

This study offers a distinctive window into how South African news publishers are managing AI content extraction, and serves as a point of departure for broader reflections on the relationship between journalism and digital platforms in the Global South.

Despite its limitations, most notably the technical constraints of robots.txt as a content protection tool and the reliance on public data, the study finds a higher AI-crawler blocking rate in South Africa than in the Brazil and Indonesia samples. However, the headline figure of 30.4% reveals a deeper structural pattern. South Africa's media sector is significantly smaller than those of Indonesia and Brazil. As confirmed by the MDPMI, the blocking rate also masks a fundamental asymmetry: the capacity to restrict AI crawler access is generally concentrated among large consolidated publishers, whilst smaller community, vernacular, and independent media outlets in South Africa lack the resources or bargaining power to implement meaningful protections.

South Africa's competition policy environment provides a stronger foundation for coordinated action than exists in most comparable markets. This foundation does not yet extend to copyright. The Constitutional Court ruling in June 2026 upheld the CAB's new fair use provisions and struck down some of its educational copying exceptions. The CAB must therefore return to Parliament. As a result, the limits of the new fair use provisions cannot yet be tested by publishers as they navigate AI content extraction.

# Next steps include:

1. **Further research:** The Media Leadership Think Tank (MLTT) at GIBS is conducting qualitative research with South African publishers across ownership tiers and outlet types to better understand the reasoning behind crawling decisions and document the practical barriers preventing smaller publishers from implementing protections.
2. **MDPMI remedy monitoring:** Together with civil society partners, the MLTT will track and report on the implementation of MDPMI's platform-specific remedies, including AI training opt-outs, publisher engagement, and capacity-building, and assess whether these remedies meaningfully shift the conditions for publishers managing AI crawler access.
3. **Copyright Amendment Bill ruling:** The MLTT will track the Bill's return to Parliament and its implications for publishers and AI content extraction.
4. **Cross-border comparison:** This report contributes to a rich, comparative dataset spanning Brazil, South Africa and Indonesia. The tricontinental CTRL+J partnership will continue to develop shared frameworks and policy recommendations grounded in Global South media conditions.

We encourage researchers in other markets to replicate this study and consult the full methodology that follows.

# Methodology

## 1.1. Data origins

This project used programmatic file verification to ascertain the existence of blocking parameters for AI agents in robots.txt files of 263 news organisations' websites from South Africa. Website data from South Africa was based on publicly available data collected and cleaned in November 2025, including membership data from the Association of Independent Publishers (AIP). The main variables considered were the name of the news organisations, their website addresses and their locations.

The methodology centred on examining these files for specific patterns and their relations to AI crawlers, based on the standard parameter Disallow:/ in robots.txt files. By consulting an open-source database of known AI crawler identifiers,[2] including those from major technology companies like Google, Meta, Amazon, Apple, Anthropic and OpenAI, the analysis identified when websites specifically attempted to signal restrictions to AI crawlers towards their content.

This aspect of the methodology was particularly relevant given the recurring concerns[3] about AI systems scraping web content without explicit permission. It is worth noting that robots.txt files are non-binding instruments that cannot technically prevent bots from collecting data, but merely indicate a site's policy towards automated content aggregation.

When examining a website, the system first ensured proper URL formatting before attempting to access its robots.txt file. The system then implemented error handling and timeout protocols to manage various real-world scenarios such as server timeouts, missing files, connection issues or DNS blocking, creating a comprehensive logging system.

To account for slow-responding servers that might need time to "wake up," the system implemented a retry mechanism with a 3-second delay when robots.txt files were not found on the first attempt. This approach significantly improved detection reliability while maintaining efficient processing times for responsive servers.

The parsing logic distinguished between restrictive directives (Disallow: / or Disallow: *) and permissive ones (Allow: directives or empty Disallow: statements), ensuring accurate classification of AI blocking policies. Sites were categorised along a spectrum from permissive to super restrictive based on the number and type of AI agents they blocked.

Websites with malformed or incomplete robots.txt files were generally classified as permissive toward AI crawlers. This included sites with empty Disallow: directives (which technically allow access) and those using only partial blocking rules that restricted specific URLs or directories while leaving the majority of their content unrestricted. In the absence of comprehensive blocking directives, these sites were effectively treated as allowing AI access to their content.

We manually checked 90 random results and were able to check that most of the results were indeed accurate. Despite these robust methodological safeguards, out of extra caution, we estimate that a 3% error rate persists due to technical factors including DNS resolution issues, server timeouts, and temporary network connectivity problems that can result in false positives or, in some cases, false negatives in robots.txt detection.

[2] List of known AI crawlers, 2025, Github repository - ai.robots.txt

[3] 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.

## 1.2 Data collection:

Once the robots.txt content was successfully retrieved, the analysis moved to pattern recognition. The system processed the file's content line by line, identifying User-agent directives and examining them against the known AI crawler database (looking for the /Disallow parameter).

This pattern matching was implemented with case-insensitive comparison to catch variations in how crawler names might be written, making the analysis less prone to redundancies. The results of this analysis were structured to provide insights about each website's stance on AI crawlers.

For each analysed website, the system determined whether a robots.txt file existed, whether it contained any blocking directives, and specifically which known AI crawlers were being blocked under the /Disallow parameter.

This structured approach to data collection and analysis made it possible to conduct fast (average of 90 minutes of code run) large-scale examinations of how different websites approached AI crawler access.

Collection and analysis were based on the R programming language.

## 1.3 Data classification:

As data was collected from the robots.txt files, the algorithm automatically made classifications based on the level of permissiveness discretely assigned to each site.

Rather than using arbitrary absolute thresholds, the system first calculates the baseline average number of AI agents blocked by sites that actually implement AI blocking policies. This average serves as the reference point for determining what constitutes normal, moderate, or excessive blocking behaviour within the specific dataset being analysed.

The classification system then creates five distinct tiers using multiples of this calculated average.

| **Classification** | **AI Agents Blocked** | **Description** |
|---|---|---|
| Permissive AI agent policy | 0 AI agents | Sites that impose no restrictions on AI crawlers |
| Somewhat restrictive | 1x to 2x average | Sites with typical or slightly above-average blocking behaviour |
| Restrictive | 2x to 5x average | Sites blocking two to five times the average |
| Very restrictive | 5x to 10x average | Sites blocking five to ten times the average |
| Super restrictive | 10x+ average | Sites with exceptionally comprehensive blocking policies |

This methodology ensures that the classification remains meaningful regardless of the absolute numbers in any given dataset, as it adapts to the actual distribution of blocking behaviours observed. For instance, if the average site blocks three AI agents, a site blocking thirty agents would be classified as super restrictive, but if the dataset average were ten agents, that same thirty-agent site would fall into the restrictive category.

## 1.4 Data aggregations

Finally, the data was aggregated in a set of indicators that could be representative of the whole dataset.

| **metric** | **value / unit** | **date_observation** |
|---|---|---|
| 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: / Use 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 Dis 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: SOUTH AFRICA Disallow: / User-agent: DuckA Disallow: / User-agent: EchoboxBot Disallow: / User-agen FacebookBot Disallow: / User-agent: Factset_spyderbot D User-agent: FirecrawlAgent Disallow: / User-agent: Frienc Disallow: / User-agent: Google-CloudVertexBot Disallow: / User-agent: Google-Extended Disallow: / User-agent: Goo Disallow: / User-agent: GoogleOther- August 2026 : / User GoogleOther-Video Disallow: / User-agent: GPTBot Disallo User-agent: iaskspider/2.0 Disallow: / User-agent: ICC-Cra Disallow: / User-agent: ImagesiftBot Disallow: / User-agen img2dataset Disallow: / User-agent: ISSCyberRiskCrawle User-agent: Kangaroo Bot Disallow: / User-agent: meta-ex