News Behind the News: Spillover Effects of AI Crawler Blocking Among News Publishers

Kanav Mor Rahul Telang

Heinz College, Carnegie Mellon University

Abstract

The emergence of generative AI has intensified long-standing tensions between content producers and platforms over the value of online content. We examine how news organizations responded to this shift by analyzing historical `robots.txt` snapshots from the Internet Archive's Wayback Machine, constructing a monthly panel of news publishers across the United States, Europe, and the Rest of the World from May 2022 through December 2025. Using interrupted time series and Callaway-Sant'Anna staggered difference-in-differences designs, we study publisher responses to the release of GPTBot in August 2023. We document three main findings. First, blocking accelerated after GPTBot's release across all three regions, with U.S. and European publishers increasing blocking at similar rates and more strongly than publishers in the Rest of the World. Second, paywalled outlets block more quickly and more strongly than free outlets. Third, blocking extended beyond AI-related crawlers: GPTBot blocking was followed by increased blocking of non-AI bots, with the strongest evidence of this spillover in the U.S. and Europe. These findings suggest that the rise of generative AI has contributed to a broader tightening of automated web access among news publishers, with implications extending beyond AI training to the governance of crawling on the web.

1 Introduction

The emergence of large language models has reignited a long-standing tension in the publishing industry: who captures the value of online content, and what recourse do producers have when new technologies shift that balance? For news organizations, this question is not new. The past two decades have seen successive waves of value displacement, from print to digital and from publishers to search engines and social platforms, that have fundamentally reshaped the economics of journalism. Declining advertising revenues have eroded newspapers' capacity to invest in journalist-intensive content (Angelucci and Cagé, 2019), while the rise of news aggregators has created complex dependencies between publishers and platforms. Calzada and Gil (2020) show that when Google News shut down in Spain following the introduction of a link fee, daily visits to Spanish news outlets fell by 8–14%, demonstrating that publishers had become substantially reliant on aggregators for traffic even as they objected to the terms of that relationship. Australia's News Media Bargaining Code, which compelled platforms to negotiate payments to news organizations, reflected a similar tension (Bossio et al., 2022). These episodes illustrate the broader conflict between those who produce information and those who profit from its distribution.

Generative AI represents a distinct escalation of this dynamic. Unlike search engines, which at minimum link users back to source material, large language models ingest training corpora in bulk and produce outputs that can substitute for, rather than complement, the original content. The scale of data collection involved is substantial: models such as GPT-3 (Brown et al., 2020) and LLaMA (Touvron et al., 2023) were trained on trillions of tokens drawn heavily from web-scraped sources, with Common Crawl-derived datasets accounting for the majority of their training corpora. The legal frameworks governing this appropriation remain unsettled. High-profile litigation, most notably by the *New York Times* against OpenAI (Grynbaum and Mac, 2023), has exposed fundamental disagreements over whether the ingestion of copyrighted material for model training constitutes fair use. Legal scholars have only begun to formalize this debate (Lucchi, 2024; Strowel, 2023). In the face of this uncertainty, publishers have been forced to act through whatever technical mechanisms are available to them.

The mechanism that has emerged as the primary instrument of response is `robots.txt`, a voluntary protocol through which website operators can signal whether specific crawlers may access site content (Koster et al., 2022). It carries no legal force; compliance depends on the good faith of the crawling party. For most of its three-decade existence, `robots.txt` operated as a background feature of web infrastructure, primarily used to prevent indexing of administrative directories or to manage server load. Its transformation into a frontline tool for asserting control over AI training data represents a striking case of institutional repurposing, what might be understood through Lessig's (1999) framework in which technical architecture

Email: kkanav@andrew.cmu.edu; rtelang@andrew.cmu.edu. No funding was required for this project. The authors declare no potential conflicts of interest.

functions as a regulatory modality when formal legal structures prove inadequate or too slow to respond. Alternative proposals for managing AI crawling have seen negligible adoption: the "noai" meta tag appeared on fewer than 0.2% of analyzed websites (Bouchaud and Ramaciotti, 2025), and the TDM Reservation Protocol has similarly seen low uptake (Dinzinger et al., 2024).

A defining moment in the rise of generative AI was the launch of ChatGPT on November 30, 2022. For publishers, however, the more actionable inflection point came with the introduction of GPTBot in August 2023. Before GPTBot, AI-related crawling was often indistinguishable from search engine crawling, and blocking general-purpose bots such as Googlebot incurred severe penalties in terms of search visibility and organic traffic. GPTBot created a separating mechanism. By publishing a distinct user-agent identifier exclusively for AI training purposes, OpenAI gave publishers the ability to block AI crawling without sacrificing search indexing. This made GPTBot's release a sharp, identifiable event around which to study how different types of organizations, with different economic exposures to AI content appropriation, responded to the same technological change.

Recent work has begun to document the broader tightening of access to web data for AI training. Longpre et al. (2024) track the rapid decline in permissive data licensing across web sources, showing that restrictions on AI data collection have accelerated sharply since 2023 and that, when Terms of Service are considered alongside `robots.txt`, the fraction of restricted content in major training datasets such as C4 rises to 45%. The resulting skew is compounded by the fact that compliance with `robots.txt` is uneven: while Liu et al. (2025) and Kim et al. (2025) find that major AI companies generally respect `robots.txt` directives, compliance varies by crawler type, and the protocol's efficacy as a protective tool remains limited.

Within this broader shift, work on publisher behavior has begun to document the descriptive landscape of blocking. Bouchaud and Ramaciotti (2025) analyze crawler restrictions across the top one million most-visited websites worldwide, demonstrating that blocking patterns vary substantially by website popularity and content type: over 25% of the top thousand websites restrict GPTBot, and among news outlets specifically, 34.2% block GPTBot, rising to 55% among outlets with high factual reporting. Zhu (2025) formalizes this pattern as adverse selection in the AI data commons. Studying more than 9,600 media sites rated by Media Bias/Fact Check, they document a steep quality-blocking gradient, showing that high-factual outlets block at nearly six times the rate of low-factual sources. They also find that publishers strategically differentiate between training and search crawlers, blocking the former at approximately 38% higher rates. Their calibrated simulations estimate that the compositional pattern of blocking has produced a 24.3% decline in the share of high-factual content in available training data.

Together, these studies establish that publishers block AI crawlers at significant rates, that this blocking is unevenly distributed, and that it is beginning to reshape the composition of AI training data. However, several questions remain open. First, existing work documents what publishers are blocking and at what rates, but the evidence is largely cross-sectional. It does not tell us how blocking behavior evolved over time in response to specific events. Second, and more fundamentally, no study has examined whether the act of blocking AI crawlers altered publishers' broader posture toward web crawling. In particular, it remains unclear whether blocking GPTBot also triggered increased blocking of non-AI crawlers that serve entirely different functions, such as SEO and web analytics tools, social media preview generators, academic and research utilities, and news monitoring agents. We refer to this possibility as the *spillover effect*.

The *spillover effect* is both empirically novel and substantively important. There are several reasons to expect that blocking GPTBot could change behavior toward unrelated crawlers. The simplest mechanism is attentional: the GPTBot debate may have prompted publishers to audit their `robots.txt` files for the first time in years, leading to the discovery and subsequent blocking of previously unknown or disregarded crawlers. A second mechanism is dispositional: the experience of perceived content appropriation may have shifted publishers toward a more defensive posture regarding all automated access, reducing their tolerance for any form of crawling that does not deliver clear, direct benefits.

Such a spillover has significant implications for the broader web ecosystem. If the response to AI crawling has generalized into a posture of broad-spectrum blocking, the consequences extend well beyond the AI training debate. They affect the infrastructure of digital transparency, academic research, media monitoring, and the web's archival record. As Hingle and Knodel (2025) argue, the current heightened attention to `robots.txt` carries implications not only for AI governance but also for the openness and accessibility of the web as a public information resource.

Using historical `robots.txt` data from the Wayback Machine, this paper makes two contributions. First, we estimate the response to the arrival of a blockable AI crawler, documenting the timing, magnitude, and heterogeneity of GPTBot blocking across U.S. and international outlets. Second, we estimate the *spillover effect* by examining whether blocking GPTBot also increased the blocking of non-AI crawlers. To do so, we use each newspaper's GPTBot blocking as the treatment date in a staggered event-study design.

2 Institutional Setting and Data

2.1 The Robots Exclusion Protocol

The Robots Exclusion Protocol is a convention, introduced by Martijn Koster in 1994, through which website operators communicate crawling permissions to automated agents. A `robots.txt` file placed at a website's root directory contains directives specifying which user-agents, identified by name, are permitted or prohibited from accessing which URL paths. For example, a directive of the form `User-agent: GPTBot` / `Disallow: /` instructs GPTBot to refrain from accessing any page on the site. The protocol was formalized as RFC 9309 by the Internet Engineering Task Force in 2022 (Koster et al., 2022), but compliance remains entirely voluntary: there is no technical enforcement mechanism, and crawlers that ignore `robots.txt` directives face no automatic sanction.

On August 7, 2023, OpenAI publicly announced GPTBot, the web crawler it uses to collect training data for its models, and simultaneously published documentation enabling website operators to block it via `robots.txt`. This announcement was notable because it gave publishers, for the first time, a named user-agent string associated with a major AI company's data collection activity and therefore a concrete target for `robots.txt` directives.

2.2 Data

We construct a panel dataset tracking the `robots.txt` configurations of news publishers over a 44-month period from May 2022 through December 2025. Our newspaper sample comprises 136 news publishers, yielding a balanced panel of 5,984 outlet-month observations. We track 50 U.S., 40 European, and 46 Rest of World (RoW) news publishers, consisting primarily of large metropolitan newspapers, broadcasters, and digital-native news websites. The data source is the Internet Archive's Wayback Machine, which periodically archives snapshots of web pages, including `robots.txt` files, and makes them accessible through its CDX Server API. For each website in our sample, we query historical `robots.txt` snapshots at monthly intervals, parse the directives contained in each snapshot, and extract the set of distinctly named user-agents subject to at least one `Disallow` directive.

Our outcome variable is the count of distinct bots blocked, that is, the number of distinct user-agents that appear with at least one `Disallow` rule in a given website's `robots.txt` file at a given point in time. This is an extensive-margin measure: it captures how many bots a site chooses to restrict, rather than the scope or stringency of each restriction. A site blocking a bot from a single subdirectory and a site blocking it from its entire domain both count as one blocked bot. We also manually decompose this count into AI-associated bots, which are linked to known AI training operations such as GPTBot, CCBot, Google-Extended, and ClaudeBot, and non-AI bots, which serve other functions such as SEO tools, social media preview generators, academic utilities, and news aggregators.

2.3 Descriptive Patterns

GPTBot blocking by News Publishers. Figure 1 displays the percentage of news publishers that block GPTBot by each calendar month. Blocking was sharply front-loaded: it began immediately following GPTBot's announcement in August 2023 and accelerated rapidly, reaching 57% of the full sample by the time *New York Times* filed a lawsuit against OpenAI in December 2023. Blocking continued at a declining rate through 2024 and 2025, ultimately reaching 81% of publishers by December 2025.

This concentrated blocking also coincided with a broader public scrutiny of AI training practices. In September 2023, reports indicated that the *New York Times* was

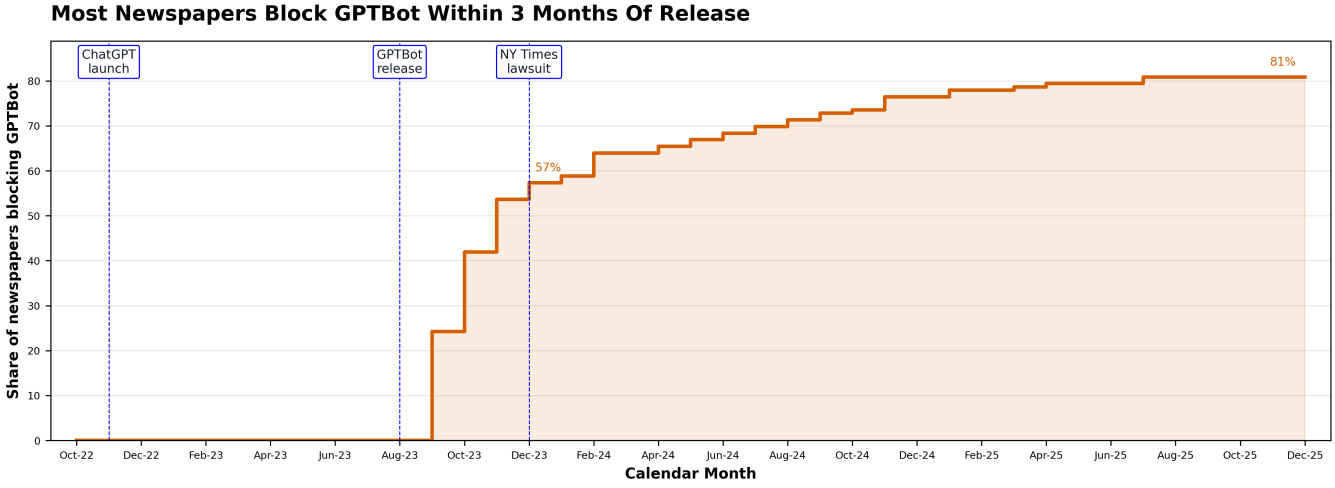


Figure 1: Percentage of newspapers in the sample blocking GPTBot

considering legal action against OpenAI, who was also pursuing licensing discussions with several publishers, including Axel Springer and the Associated Press. In December 2023, the *New York Times* filed a lawsuit against OpenAI, alleging unauthorized use of its content for model training (Grynbaum and Mac, 2023). These developments may have increased the salience of decisions about crawler access during this period, even though the descriptive evidence here does not isolate their separate effects.

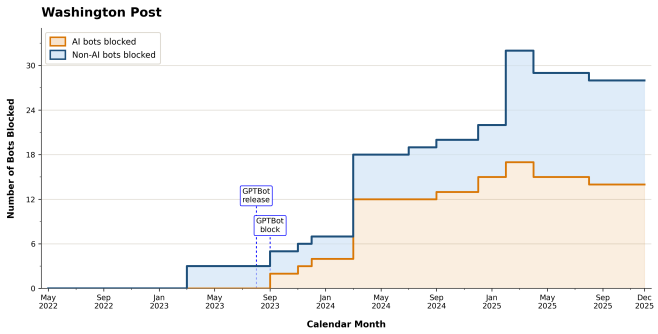


Figure 2: AI versus non-AI bots blocked by WPost

Washington Post as a case study. Figure 2 presents the blocking trajectory of *Washington Post*, which is among the more aggressive responders in our sample. While Figure 1 establishes that GPTBot blocking was rapid and widespread, it does not reveal what else publishers changed in their `robots.txt` files at the same time. Figure 2 decomposes total bots blocked into AI and non-AI bots for *Washington Post*. Blocking rose in a stepwise pattern, increasing from 18 bots in March 2024 to 32 by February 2025. Critically, the non-AI count grew in tandem with the AI count, ultimately accounting for roughly half of all blocked bots by the end of 2025.

U.S. Publishers' Blocking by Crawler Type. GPTBot blocking often coincided with, or was followed by, increased blocking of crawlers with no known AI training function. These include SEO and analytics tools such as AhrefsBot and SemrushBot, social media preview generators such as FacebookBot, academic and research utilities such as TurnitinBot, and news monitoring agents such as NewsNow. Many of these crawlers predate GPTBot by more than a decade. Their blocking by news websites after 2023, despite years of prior coexistence, suggests that the response to AI crawling was not narrowly targeted at AI-specific agents but extended to a broader reassessment of automated access.

Figure 3 depicts the blocking by an average US publisher, disaggregating blocking by crawler category. Prior to GPTBot's release, the average U.S. publisher blocked approximately one bot. By December 2025, this figure had risen to approximately 17, representing a sustained increase over the intervening 28 months. AI-related crawlers account for the largest single category throughout the post-GPTBot period, but the non-AI categories collectively account for a substantial and growing share. The figure also reveals a visible acceleration in blocking activity after the launch of AI Overviews in Google Search in May 2024. The formal estimation in Section 3 tests whether this increase in non-AI blocking is statistically attributable to the GPTBot event or instead reflects independent trends.

Bot Blocking Across Regions. Figure 4 compares crawler-blocking trajectories across U.S., European, and Rest-of-World publishers. Overall blocking is largest among U.S. and European publishers and seems to increase only after GPTBot's release. The bottom row focuses on non-AI bots and centers time around each publisher's first GPTBot block month. It examines whether blocking GPTBot coincided with a broader tightening toward other crawlers. It suggests that non-AI blocking rises after GPTBot blocking in the U.S. and Europe, while the rest-of-world blocking remains flat.

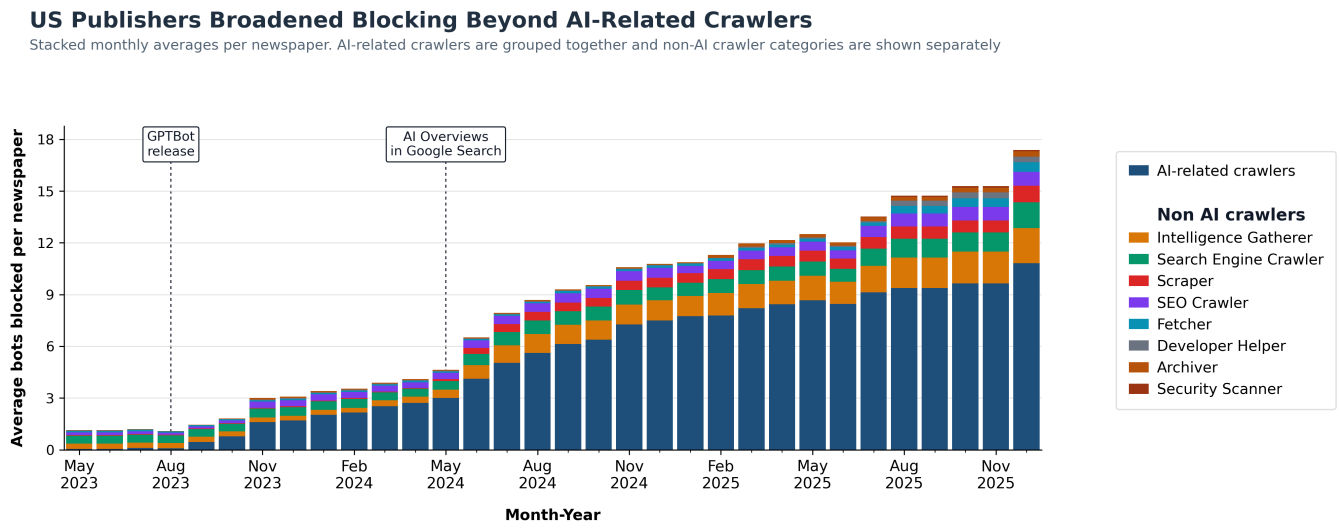


Figure 3: Average bots blocked per U.S. publisher, by crawler category

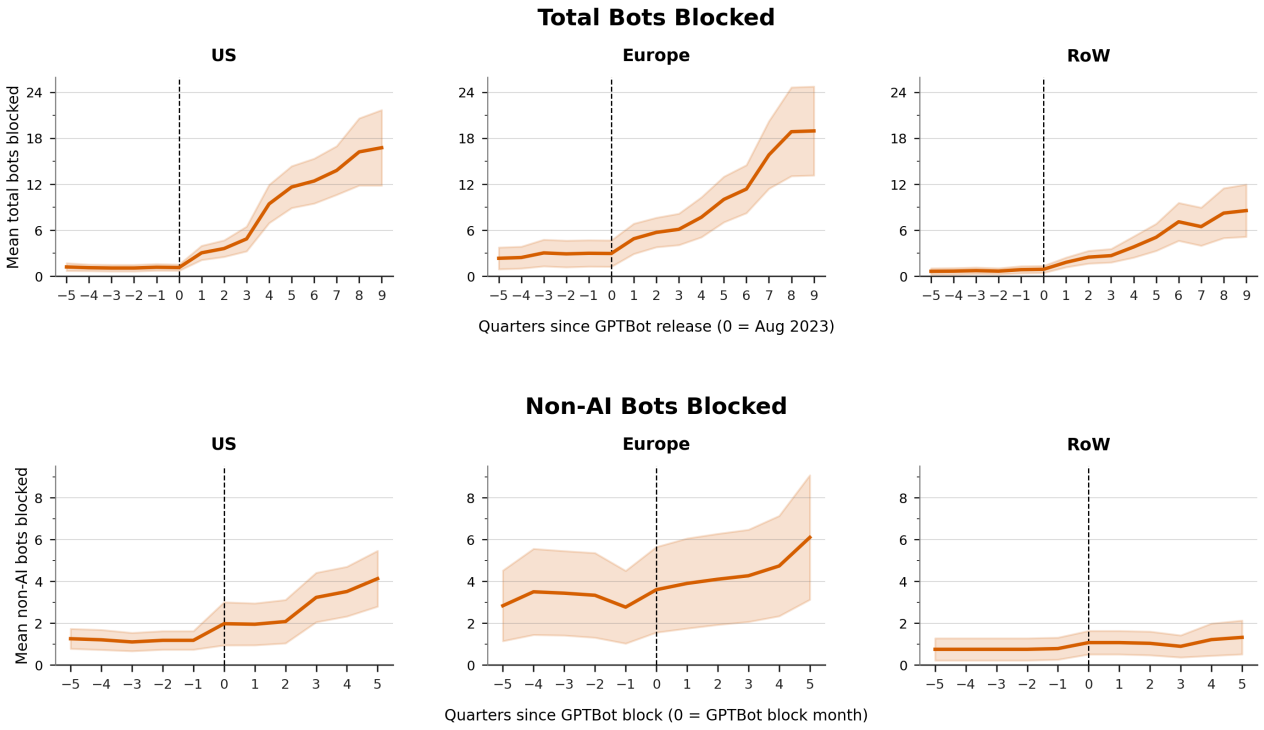


Figure 4: Average bots blocked by news publishers from each region

3 Empirical Strategy

Estimating news publishers' response to GPTBot's release presents a fundamental challenge: the launch of generative AI was global and simultaneously affected all online content producers. There was no natural population of news publishers that remained unexposed to the launch of generative AI. In the absence of a clean counterfactual, we use an interrupted time series (ITS) design to estimate aggregate blocking by news publishers, which tests whether publisher blocking exhibits structural breaks at shocks like GPTBot's release and the later launch of AI Overviews in Google Search. To estimate spillover, we use a Callaway-Sant'Anna staggered difference-in-differences design that uses the month of each publisher's first GPTBot block as its treatment date and compares treated publishers to not-yet-treated publishers in the same region. The first design characterizes the aggregate tightening of `robots.txt` restrictions; the second tests whether GPTBot blocking was followed by broader restrictions on crawlers with no known AI training function.

3.1 Total Bots Blocked: Interrupted Time Series

The ITS design models the time path of total bot blocking among news publishers as a piecewise-linear function with structural breaks at two intervention dates: the release of GPTBot in August 2023 and the launch of AI Overviews in Google Search in May 2024. At each intervention, the model allows for two changes: a discrete level shift - an immediate change in the number of bots blocked, and a slope change - an acceleration or deceleration in the rate of blocking over subsequent months. The second break is included to distinguish the initial response to GPTBot from a later but distinct AI-related shock that may have renewed or heightened publisher concern about the use of their content. The full model specification is reported in Appendix A, where entity fixed effects absorb time-invariant differences across outlets. For robustness, we also run placebo tests that re-estimate the model on pre-intervention data using fictitious break dates.

3.2 Non-AI Bots Blocked: Staggered Difference-in-Differences

The ITS design captures common changes in total bot blocking, but the spillover question is different. We want to know whether a publisher's decision to block GPTBot is followed by additional blocking of non-AI crawlers. For this outcome, treatment timing varies across publishers because outlets adopted GPTBot blocking in different months. We use this staggered adoption structure to estimate region-specific Callaway-Sant'Anna difference-in-differences models.

Let G_i^r denote the first month in which publisher i in region r blocks GPTBot, where $r \in \{\text{US}, \text{Europe}, \text{RoW}\}$. The outcome, Y_{it}, is the number of non-AI bots blocked by publisher i in month t. For each region separately, we estimate group-time average treatment effects, where

g indexes the treatment cohort and t indexes calendar month:

$$ATT^r(g,t) = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid G_i^r = g], \qquad t \geq g.$$

Treated publishers are compared only to publishers in the same region that have not yet blocked GPTBot by month t. Let $k = t - g$ denote event time relative to the first GPTBot block. We aggregate the estimated group-time effects by monthly event time:

$$\theta_k^r = \sum_g w_{g,k}^r ATT^r(g, g+k)$$

Here, θ_k^r is the region-specific dynamic treatment effect at event time k, and $w_{g,k}^r$ are the cohort-specific aggregation weights within region r. This design identifies the spillover effect without relying on institutionally distinct website categories as a comparison group.

4 Results

This section presents three main findings. First, total bot blocking by news publishers is essentially flat prior to GPTBot's release and increases only after August 2023, with additional tightening after the launch of Google's AI Overviews. Second, paywalled outlets respond more strongly than free outlets. Third, GPTBot blocking is followed by increased blocking of non-AI crawlers, with the strongest evidence of spillover in the U.S. and Europe.

4.1 Total Bots Blocked

Appendix Table A.1 presents the ITS estimates, and Figure 5 plots the corresponding segmented fits by region. In the pre-GPTBot period, the estimated monthly trend is close to zero across all three regions, indicating that `robots.txt` configurations were largely stable prior to August 2023.

After GPTBot's release, blocking begins to accelerate. In the U.S., the estimated slope rises by 0.398 bots per month ($p < 0.01$). Europe follows a statistically similar post-GPTBot trajectory, while RoW publishers also increase blocking but at a significantly smaller rate, roughly half the U.S. pace. The fitted paths therefore suggest a gradual diffusion process in which publishers tightened `robots.txt` restrictions over successive months.

The launch of AI Overviews in Google Search in May 2024 is associated with a further steepening of the trajectory, especially in the U.S. and Europe. Taken together, the ITS estimates indicate sustained tightening in publisher blocking after August 2023, with a further intensification after May 2024 concentrated primarily in the U.S. and Europe.

As a complementary analysis, Table 1 reports total-bots blocked in the post period after controlling for entity fixed effects. The estimates indicate a large increase in average aggregate blocking after GPTBot's release: 8.5 additional bots in the U.S., 7.8 in Europe, and 4.1 in RoW. The regional ordering mirrors the ITS results, with the strongest response in the U.S. and Europe and a smaller response in RoW.

$$\textit{Total Bots Blocked}_{it} = \alpha_i + \beta \textit{Post}_t + \varepsilon_{it}$$

Table 1: Total Bots Blocked

	US	Europe	RoW
Post	8.477***	7.781***	4.073***
	(1.052)	(1.083)	(0.764)
Entity FE	YES	YES	YES

Notes: Rows report coefficients from separate region-specific models for total bots blocked. Standard errors are in parentheses and are clustered at the entity level.

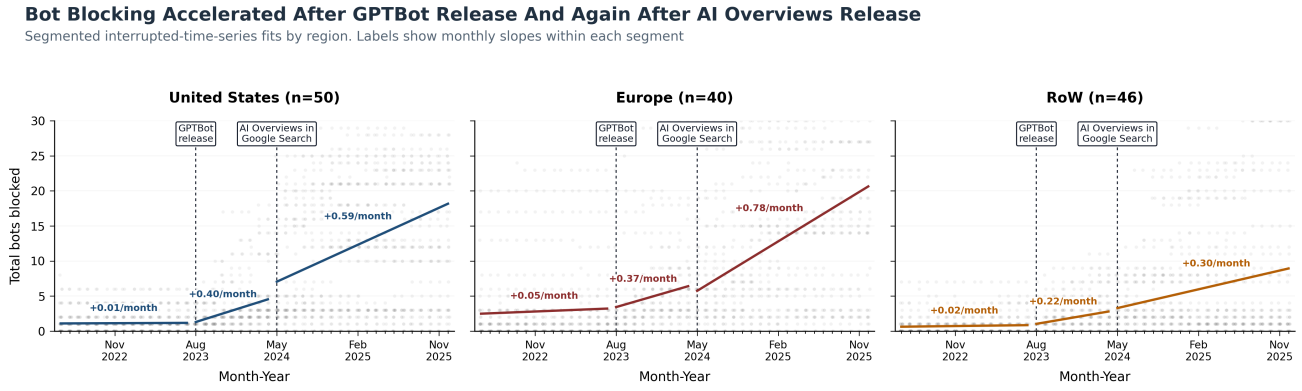


Figure 5: Interrupted time series for total bots blocked by news publishers

4.2 Paywall Status Heterogeneity

We next examine whether the blocking response differed by pre-treatment paywall status. Paywall status is measured as of November 2022, before ChatGPT's launch, and we find that 54 of the 136 publishers in our sample were paywalled, while 82 were free outlets. We do not study the heterogeneity between hard, metered, and freemium paywall models - all three are simply counted as paywalled in the analysis. If the perceived costs of AI-mediated appropriation were greater for outlets that rely more heavily on controlling access to content, one would expect stronger blocking among paywalled publishers.

Figure 6 depicts blocking behavior by paywall status. Paywalled outlets were more likely to ever block GPTBot: 91% of paywalled publishers blocked GPTBot, compared with 74% of free outlets. At the peak, the average number of blocked bots reached 24 for paywalled outlets versus 12 for free ones. Paywall status is associated with a 15.3 percentage-point higher probability of ever blocking GPTBot (Table C.1). Similar patterns appear for total bots blocked. The average news publisher increased blocking by about 5.3 bots after GPTBot's release, while paywalled outlets exhibit an additional increase of about 3.8 bots above this baseline (Table C.2). The blocking response is therefore substantially larger among outlets with pre-treatment paywalls, which likely had stronger ex ante incentives to control access to their content.

4.3 Non-AI Bots Blocked

We next examine whether GPTBot blocking was followed by broader restrictions on non-AI crawlers. Table 2 reports the overall ATT from region-specific Callaway-Sant'Anna staggered DiD models, using not-yet-treated publishers in the same region as controls. The estimated effect is positive and statistically significant in the U.S. and Europe: GPTBot blocking is followed by 0.95 additional non-AI bots blocked in the U.S. and 2.50 in Europe. Although the European point estimate is larger, the confidence intervals for the U.S. and Europe overlap substantially. When we formally test whether the two coefficients differ, we cannot reject equality. The RoW estimate is negative and statistically indistinguishable from zero.

Table 2: Overall ATT: Non-AI Bots Blocked

	US	Europe	RoW
Overall ATT	0.951***	2.502**	-0.257
	(0.275)	(1.119)	(0.278)
Time FE	YES	YES	YES
Entity FE	YES	YES	YES
Counterfactual	Not-yet-treated	Not-yet-treated	Not-yet-treated

Notes: Rows report simple overall ATT estimates from region-specific Callaway-Sant'Anna models. *, **, and *** denote $p < 0.10$, $p < 0.05$, and $p < 0.01$, respectively.

The dynamic estimates are depicted in Figure 7. First, the pre-treatment estimates are close to zero in all three regions, indicating no evidence that publishers who eventually blocked GPTBot were already changing their non-AI blocking behavior beforehand. Second, the spillover in the U.S. and Europe builds over time rather than appearing immediately: post-treatment effects become statistically significant only from the third quarter onward. In RoW, the dynamic estimates remain small and imprecise, providing little evidence of comparable spillover. Detailed event-time estimates are reported in Appendix Table B.1.

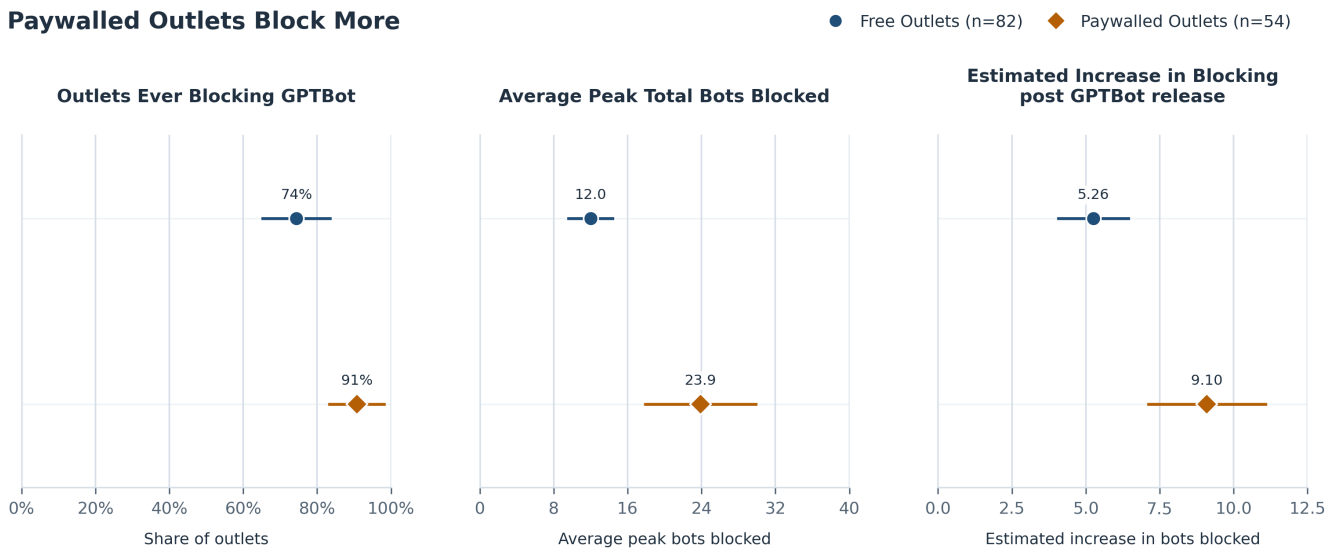


Figure 6: Paywall heterogeneity in total bots blocked

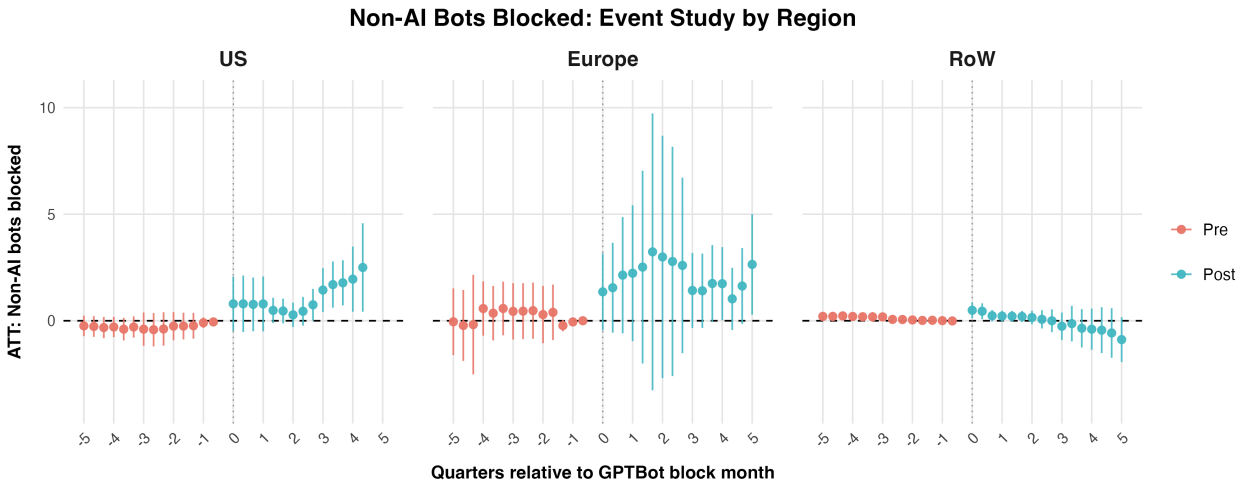


Figure 7: Callaway-Sant'Anna event study for non-AI bots blocked

5 Discussion

5.1 Interpreting the Spillover

The paper's main substantive result is not only that news publishers blocked GPTBot, but that blocking extended to non-AI crawlers as well. The timing of that extension is informative. In the dynamic estimates, non-AI blocking does not rise immediately at GPTBot's introduction; instead, it emerges with a lag and builds over subsequent quarters. A plausible interpretation is that GPTBot served as a trigger for publishers to revisit their broader crawling policies.

One channel is *attentional audit*. Once publishers incurred the fixed cost of reviewing and updating their `robots.txt` files, the marginal cost of adding additional crawlers to a deny list was low. In that setting, a narrowly targeted response to AI crawling could readily broaden into a wider tightening of automated access.

Another channel is *defensive posturing*. The boundary between AI-related and non-AI crawling is not always sharp, either because some services combine multiple functions or because publishers may be uncertain about how particular crawlers use their content. Under such uncertainty, and in the presence of asymmetric downside from unauthorized reuse, overblocking may be a rational defensive strategy. The observed spillover may therefore reflect broad blocking under uncertainty, using a coarse control instrument - robots.txt.

5.2 Regional and Paywall Asymmetries

Response to GPTBot was shaped by both - the external environment and publisher-specific incentives. Regionally, the effects are strongest in the United States and Europe, and weaker in the Rest of World. One interpretation is that GPTBot's release interacted with a policy and media environment in which concerns about AI appropriation were already especially salient in the U.S. and Europe, owing to litigation, licensing negotiations, and broader public debate. The weaker average response in RoW may partly reflect less intense salience, but also the fact that this category pools a more heterogeneous set of markets and institutional settings. So, that should be interpreted with care.

The paywall results point to another source of heterogeneity. Publishers with pre-treatment paywalls were more likely to block GPTBot and blocked more bots overall than free outlets. As their business model depends more directly on limiting access to proprietary content, they perceive higher risk that AI systems would substitute for, summarize, or otherwise appropriate the value of original reporting.

Taken together, these heterogeneity patterns suggest that GPTBot did not induce a uniform technological response. Rather, it activated different levels of concern across publishers, with the strongest reactions occurring where AI-related appropriation was likely to be perceived as both salient and economically consequential.

5.3 Implications for Web Governance

The broader implication of these findings is that `robots.txt` is being used to address a problem for which it was not originally designed. The protocol operates at the level of crawler access: it can allow or deny entry, but it cannot readily distinguish among downstream uses of content, nor can it impose legally binding conditions on use after access is allowed. When publishers rely on it to respond to AI training concerns, broader restrictions on

other forms of crawling become more likely.

That mismatch matters because many of the non-AI crawlers being blocked do not perform functions that are straightforwardly adverse to publishers. Some support content monitoring, plagiarism detection, archiving, news aggregation, etc. Restricting such crawlers may be privately rational for publishers seeking greater control over access to their content, but it can also limit services that have historically contributed to discoverability, accountability, and interoperability on the web. In this sense, the spread of GPTBot blocking appears to reflect not only a dispute over AI training data, but also a broader tightening of the terms under which automated access to web content is permitted.

More generally, we see a governance problem: the underlying conflict concerns particular uses of content, while the available technical control is a broad crawler-level exclusion rule. As long as that mismatch persists, publishers may continue to respond in ways that are broader than the original concern. A more durable resolution would likely require governance mechanisms that allow publishers to distinguish more precisely among access, use, and permission than the current `robots.txt` framework permits.

6 Limitations

Few limitations should be noted. First, our outcome variable captures the extensive margin of blocking - the count of distinct bots restricted, but not the intensive margin - the scope of restriction for a bot in the `robots.txt` file. A newspaper that blocks a crawler from a single subdirectory and one that blocks it from the entire domain are treated identically. As a result, the measure does not fully capture heterogeneity in the stringency of publisher responses.

Second, `robots.txt` directives are requests, not enforceable restrictions. Our data therefore measure what publishers ask crawlers to do, not what crawlers actually do. The efficacy of these directives as a protective mechanism depends on crawler compliance, which, as prior work has shown, is variable. Accordingly, our findings speak to changes in publishers' recorded restrictions on automated access, not to the actual quantity of scraping prevented. Also, we do not observe how `robots.txt` changes were made within organizations - whether they were driven by legal, editorial, technical, or commercial actors, or whether they reflected deliberate policy choices rather than routine administrative updates.

7 Conclusion

This paper examines how news publishers changed their `robots.txt` policies after the release of GPTBot. Using historical snapshots from the Wayback Machine, we show that publisher blocking was largely stable before August 2023 and increased sharply afterward. The response was especially pronounced among U.S. and European publishers, and among outlets with pre-treatment paywalls. These patterns suggest that GPTBot's release pushed publishers to pay closer attention to crawler access, particularly those with stronger incentives to control access to proprietary content.

The main finding, however, is not limited to GPTBot or to AI crawlers. Publishers that blocked GPTBot also increased blocking of non-AI crawlers, with the strongest evidence in the U.S. and Europe. This spillover indicates that the response to AI crawling was not always narrowly targeted. Instead, GPTBot appears to have triggered a broader reassessment of crawler access, extending restrictions to tools used for search visibility, monitoring, aggregation, archiving, and other non-AI functions.

Taken together, these results point to a mismatch between the concern publishers face and the tool available to them. The concern is not automated access alone, but how accessed content may be used downstream. Yet `robots.txt` operates at the crawler level: it can allow or deny an agent, but it cannot specify whether access is permitted for indexing, monitoring, archiving, AI training, or other uses. As a result, attempts to control AI training can spill over into broader restrictions on web crawling. Reducing overblocking will require mechanisms that let publishers express permissions by use, and not only by crawler identity.

References

Angelucci, Charles, and Julia Cagé. 2019. "Newspapers in Times of Low Advertising Revenues." *American Economic Journal: Microeconomics* 11(3): 319–364. doi:10.1257/mic.20170306.

Bossio, Diana, Terry Flew, James Meese, Tama Leaver, and Belinda Barnet. 2022. "Australia's News Media Bargaining Code and the Global Turn Towards Platform Regulation." *Policy & Internet* 14(1): 136–150. doi:10.1002/poi3.284.

Bouchaud, Paul, and Pedro Ramaciotti. 2025. "Web Crawler Restrictions, AI Training Datasets & Political Biases." arXiv:2510.09031.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, Vol. 33, 1877–1901.

Calzada, Joan, and Ricard Gil. 2020. "What Do News Aggregators Do? Evidence from Google News in Spain and Germany." *Marketing Science* 39(1): 134–167. doi:10.1287/mksc.2019.1150.

Dinzinger, Michael, Florian Heß, and Michael Granitzer. 2024. "A Survey of Web Content Control for Generative AI." In *Proceedings of the Workshop on Open Web Search (WOWS@ECIR)*, CEUR Workshop Proceedings, Vol. 3689, 2–13.

Grynbaum, Michael M., and Ryan Mac. 2023. "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work." *The New York Times*, December 27.

Hingle, Audrey, and Mallory Knodel. 2025. "Robots.txt Is Having a Moment: Here's Why We Should Care." *Tech Policy Press*, April 3.

Kim, Taein, Karstan Bock, Claire Luo, Amanda Liswood, Chloe Poroslay, and Emily Wenger. 2025. "Scrapers Selectively Respect robots.txt Directives: Evidence from a Large-Scale Empirical Study." In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25)*. doi:10.1145/3730567.3764471.

Koster, Martijn, Gary Illyes, Henner Zeller, and Lizzi Sassman. 2022. *Robots Exclusion Protocol (RFC 9309)*. Internet Engineering Task Force / RFC Editor. doi:10.17487/RFC9309.

Lessig, Lawrence. 1999. *Code and Other Laws of Cyberspace*. Basic Books.

Liu, Enze, Elisa Luo, Shawn Shan, Geoffrey M. Voelker, Ben Y. Zhao, and Stefan Savage. 2025. "Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers." In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25)*. doi:10.1145/3730567.3732913.

Longpre, Shayne, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, et al. 2024. "Consent in Crisis: The Rapid Decline of the AI Data Commons." In *Advances in Neural Information Processing Systems*, Vol. 37, 108042–108087. doi:10.52202/079017-3431.

Lucchi, Nicola. 2024. "ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems." *European Journal of Risk Regulation* 15(Special Issue 3): 602–624. doi:10.1017/err.2023.59.

OpenAI. n.d. "Overview of OpenAI Crawlers." OpenAI developer documentation, accessed February 23, 2026.

Strowel, Alain. 2023. "ChatGPT and Generative AI Tools: Theft of Intellectual Labor?" *IIC – International Review of Intellectual Property and Competition Law* 54(4): 491–494. doi:10.1007/s40319-023-01321-y.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, et al. 2023. "LLaMA: Open and Efficient Foundation Language Models." arXiv:2302.13971.

Zhu, K. (2026). Adverse selection in the AI data commons. SSRN. doi:10.2139/ssrn.6438640

Appendix

A Interrupted Time Series Regression

This appendix reports the interrupted time series specification used in Section 4.1, the corresponding coefficient estimates, and placebo break tests based on pre-period data only. Table A.1 reports the full coefficient vector for the segmented specification, and Table A.2 reports placebo breaks at alternative pre-period cutoffs.

$$
\begin{aligned}
Y_{irt} = \alpha_i + \beta_1 t
&+ \sum_{g \in \mathcal{G}} \beta_g (D_{gr} \times t) \\
&+ \sum_{m \in \mathcal{M}} \Big[\gamma_m Post_t^m + \kappa_m TimeSince_t^m \\
&\qquad + \sum_{g \in \mathcal{G}} \big(\delta_{mg} (D_{gr} \times Post_t^m) \\
&\qquad\qquad + \phi_{mg} (D_{gr} \times TimeSince_t^m) \big) \Big] \\
&+ \varepsilon_{irt}
\end{aligned}
$$

Definitions. Y_{irt} is the total number of bots blocked for publisher i in region r and month t, where t is measured in months over the sample period. α_i denotes entity fixed effects. $\mathcal{G} = \{\text{Europe}, \text{RoW}\}$, with the U.S. as the reference region, and $D_{gr} = \mathbf{1}\{r = g\}$. $\mathcal{M} = \{\text{GPTBot}, \text{AIO}\}$, where GPTBot refers to August 2023 and AIO refers to the May 2024 launch of AI Overviews in Google Search. For each intervention m, $Post_t^m$ equals 1 in months on or after the intervention and 0 otherwise, while $TimeSince_t^m$ counts months since the intervention and equals 0 beforehand. β_1 is the U.S. baseline monthly trend, and β_g gives region-specific deviations from that trend. γ_m and κ_m capture the U.S. level and slope changes at intervention m, while δ_{mg} and ϕ_{mg} capture Europe/RoW deviations from those level and slope changes. Region main effects are absorbed by the entity fixed effects. Standard errors are clustered at the entity level.

Table A.1: ITS Regression

	Estimate
Time trend (US baseline)	0.006
	(0.008)
Europe diff. pre-trend	0.046
	(0.044)
RoW diff. pre-trend	0.010
	(0.012)
GPTBot level shift (US)	0.124
	(0.162)
GPTBot slope change (US)	0.398***
	(0.077)
AI Overviews level shift (US)	2.119**
	(1.011)
AI Overviews slope change (US)	0.182
	(0.138)
Europe diff. GPTBot level	0.039
	(0.312)
RoW diff. GPTBot level	0.045
	(0.202)
Europe diff. AIO level	-3.156**
	(1.369)
RoW diff. AIO level	-1.845*
	(1.111)
Europe diff. GPTBot slope	-0.077
	(0.144)
RoW diff. GPTBot slope	-0.195**
	(0.095)
Europe diff. AIO slope	0.230
	(0.245)
RoW diff. AIO slope	-0.104
	(0.160)
Entity FE	YES
Region-specific pre-trends	YES

Notes: Standard errors in parentheses. Standard errors are clustered at the entity level. Entity fixed effects are included. Region-specific baseline pre-trends are allowed. The U.S. is the omitted region for interaction terms.

Table A.2: ITS Placebo Cutoff Tests

Placebo cutoff	Break type	Estimate	SE	p-value
Aug 2022 (-12)	Level shift	0.071	(0.134)	0.5969
	Slope change	0.021	(0.032)	0.5054
Nov 2022 (-9)	Level shift	0.017	(0.048)	0.7176
	Slope change	-0.008	(0.041)	0.8412
Feb 2023 (-6)	Level shift	-0.049	(0.071)	0.4912
	Slope change	0.000	(0.031)	0.9982
May 2023 (-3)	Level shift	0.004	(0.079)	0.9563
	Slope change	-0.016	(0.027)	0.5420

Notes: Each placebo cutoff is re-estimated on pre-period data only using a fake GPTBot intervention date. The table reports both placebo level shifts and placebo slope changes with entity-clustered standard errors. Significant placebo breaks in either dimension would weaken the August 2023 break design.

B Non-AI Bots Staggered Difference-in-Differences

This appendix reports the detailed dynamic Callaway-Sant'Anna estimates underlying the non-AI spillover results in Section 4.3. The overall ATT is reported in Table 2; Table B.1 reports the event-time estimates at relative-quarter marks.

Table B.1: Dynamic ATT: Non-AI Bots Blocked

	Region-specific dynamic ATT		
Event quarter q	US	Europe	RoW
$q = -3$	-0.394	0.441	0.181*
	(0.400)	(0.675)	(0.103)
$q = -2$	-0.252	0.290	0.038
	(0.339)	(0.686)	(0.076)
$q = 0$	0.795	1.357	0.491**
	(0.653)	(0.922)	(0.191)
$q = 1$	0.789	2.230	0.217*
	(0.657)	(1.627)	(0.129)
$q = 2$	0.278	2.993	0.152
	(0.293)	(2.903)	(0.162)
$q = 3$	1.441***	1.420	-0.259
	(0.523)	(0.899)	(0.329)
$q = 4$	1.952**	1.740**	-0.396
	(0.780)	(0.876)	(0.493)
$q = 5$		2.643**	-0.884*
		(1.203)	(0.536)
Time FE	YES	YES	YES
Entity FE	YES	YES	YES

Notes: Rows report the subset of region-specific Callaway-Sant'Anna monthly dynamic ATT estimates that fall on relative-quarter marks, where q indexes relative quarter. Regions are estimated separately using not-yet-treated publishers as the counterfactual group. Blank cells indicate unavailable relative-quarter estimates. Standard errors are in parentheses. *, **, and *** denote $p < 0.10$, $p < 0.05$, and $p < 0.01$, respectively.

C Paywall Heterogeneity

This appendix reports supplemental paywall heterogeneity estimates underlying Section 4.2. Table C.1 shows cross-sectional associations between pre-treatment paywall status and eventual GPTBot blocking and peak total bots blocked, while Table C.2 reports the pooled post-GPTBot heterogeneity estimate.

Table C.1: Paywall Heterogeneity: Cross-Sectional

	Ever Blocked GPTBot	Max Total Bots
Intercept	0.799***	15.085***
	(0.068)	(2.964)
`isPaywalled`	0.153**	10.631***
	(0.064)	(3.286)
RoW	-0.081	-7.850**
	(0.082)	(3.277)
US	-0.063	0.210
	(0.078)	(4.117)
Region FE	YES	YES

Notes: Standard errors in parentheses. Column (1) is estimated as a linear probability model. Standard errors are HC1 heteroskedasticity-robust. Europe is the omitted region category. The sample is one row per news outlet.

Table C.2: Paywall Heterogeneity: Pooled DID

	Total Bots
Post-GPTBot	5.258***
	(0.628)
Post-GPTBot × `isPaywalled`	3.839***
	(1.210)
Entity FE	YES

Notes: Standard errors in parentheses. *, **, and *** denote $p < 0.10$, $p < 0.05$, and $p < 0.01$, respectively. Standard errors are clustered at the entity level.

D Hazard Model of Time to Blocking

This appendix reports supplemental Cox proportional hazards models for the timing of blocking decisions. Unlike the main panel regressions, which study changes in blocking levels, these models examine time to first blocking event.

Table D.1 reports time from GPTBot's release to first GPTBot block across platform types. Relative to U.S. news outlets, video platforms exhibit a significantly lower hazard of adoption, while differences for European, RoW, and social-platform groups are not statistically significant.

Table D.1: Hazard Model: GPTBot Blocking

	Coefficient	Hazard ratio
EUR News	-0.029	0.972
	(0.237)	
RoW News	-0.384	0.681
	(0.238)	
Video Platforms	-1.377**	0.252
	(0.600)	
Social Networks	-0.599	0.549
	(0.599)	
Concordance	0.583	
Baseline group	US Outlets	

Notes: Entries report log hazard coefficients with standard errors in parentheses. Hazard ratios are shown in the right column. The outcome is months from the August 2023 GPTBot release to first GPTBot block; units without an event by December 2024 are right-censored.

Table D.2 reports time from first GPTBot block to first additional non-AI block among outlets that blocked GPTBot by November 2024. Relative to U.S. outlets, RoW outlets exhibit a significantly lower hazard of additional non-AI blocking, while the European difference is negative but not statistically significant.

Table D.2: Hazard Model: Additional Non-AI Blocking

	Coefficient	Hazard ratio
EUR outlets	-0.210	0.811
	(0.365)	
RoW outlets	-1.439***	0.237
	(0.553)	
Concordance	0.646	
Baseline group	US outlets	

Notes: Entries report log hazard coefficients with standard errors in parentheses. Hazard ratios are shown in the right column. The sample includes outlets that first blocked GPTBot by November 2024, leaving at least one month of follow-up. The outcome is months from first GPTBot block to first additional non-AI block; outlets without an event by December 2024 are right-censored.