Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers

Enze Liu*
e7liu@ucsd.edu
UC San Diego
La Jolla, CA, USA

Elisa Luo*
e4luo@ucsd.edu
UC San Diego
La Jolla, CA, USA

Shawn Shan
shansixiong@cs.uchicago.edu
University of Chicago
Chicago, IL, USA

Geoffrey M. Voelker
voelker@ucsd.edu
UC San Diego
La Jolla, CA, USA

Ben Y. Zhao
ravenben@cs.uchicago.edu
University of Chicago
Chicago, IL, USA

Stefan Savage
ssavage@ucsd.edu
UC San Diego
La Jolla, CA, USA

Abstract

The success of generative AI relies heavily on training on data scraped through extensive crawling of the Internet, a practice that has raised significant copyright, privacy, and ethical concerns. While few measures are designed to resist a resource-rich adversary determined to scrape a site, crawlers can be impacted by a range of existing tools such as robots.txt, NoAI meta tags, and active crawler blocking by reverse proxies.

In this work, we seek to understand the ability and efficacy of today's networking tools to protect content creators against AI-related crawling. For targeted populations like human artists, do they have the technical knowledge and agency to utilize crawler-blocking tools such as robots.txt, and can such tools be effective? Using large scale measurements and a targeted user study of 203 professional artists, we find strong demand for tools like robots.txt, but significantly constrained by critical hurdles in technical awareness, agency in deploying them, and limited efficacy against unresponsive crawlers. We further test and evaluate network level crawler blockers provided by reverse proxies. Despite relatively limited deployment today, they offer stronger protections against AI crawlers, but still come with their own set of limitations.

CCS Concepts

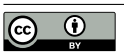
• **Information systems** → **Web crawling**.

Keywords

Robots.txt; AI Crawlers; Web Content Control; Content Creators;

ACM Reference Format:
Enze Liu, 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), October 28–31, 2025, Madison, WI, USA.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3730567.3732913

*Equal contribution.



ACM ISBN 979-8-4007-1860-1/2025/10
https://doi.org/10.1145/3730567.3732913

1 Introduction

The success of generative AI relies heavily on training on data scraped through extensive crawling of the Internet, a practice that has raised significant copyright, privacy, and ethical concerns. Today, AI model trainers have unleashed large numbers of data crawlers on the Internet. By many reports, these crawlers now dwarf the volume of human traffic on the Internet, partly because human users consume content at a much lower rate than crawlers. For example, analysis by Akamai and Imperva suggest that roughly 50–70% of website traffic is due to automated crawlers [48, 109]. Other anecdotal evidence suggests that AI crawlers are effectively producing DDoS attacks on smaller websites [25, 26].

While Internet crawling is well-studied, the widespread adoption of generative AI and its intensive data scraping has significantly changed the landscape. Data creators and hosting platforms, who were generally ambivalent about having their content crawled in the past, are now raising serious concerns about AI-related crawling, particularly regarding copyright, privacy, and ethical practices. Indeed, these concerns have manifested in over thirty ongoing copyright lawsuits [68, 69, 112], multiple data strikes [34, 118], and a surge in the adoption of anti-crawling tools [70].

Given this new tension between AI training companies seeking training data and content creators who consider unauthorized AI training an existential threat to their livelihoods [31], a natural question arises: *What tools, if any, can content creators use to prevent their content from being crawled for AI training?* Answering this question requires a more thorough understanding of the needs of content creators; their awareness of, accessibility to, and agency over anti-crawling mechanisms; and ultimately, the availability and efficacy of current tools.

This paper presents our efforts to address these issues from several complementary perspectives. In terms of representative content creators, we focus on visual artists as the most vulnerable population being targeted by AI crawlers. In terms of anti-crawling mechanisms, we focus on two tools at different ends of the spectrum. The most prominent and popular tool is robots.txt, a voluntary (and non-enforceable) protocol that enables site owners to specify crawling restrictions. We also consider crawler blocking by reverse proxies (e.g., Cloudflare), an active approach that enforces blocking but has seen limited deployment.

We begin with a longitudinal analysis of robots.txt files across the Web. Utilizing data from Common Crawl [24], we analyze the

inclusion of directives that specifically target AI crawlers over time. This effort serves as broader context on how the arrival of AI crawlers has changed views across the Web towards crawling. We then turn our attention to visual artists, and perform a user study to understand their attitudes towards AI crawlers, and their awareness of and accessibility to defensive tools like robots.txt. We complement these results with measurements of 1100+ professional artist websites to examine the hosting services artists use and levels of control these services provide. Next, we use sites under our control to determine which AI crawlers respect robots.txt. Finally, we consider active crawler blocking techniques, and measure their deployment as well as their efficacy across different AI crawlers.

Results from our study highlight critical hurdles that limit or prevent the effective utilization of protective tools by individual creators, leaving these key stakeholders in the data ecosystem vulnerable and often unable to safeguard their work from unauthorized AI-driven use. More specifically, our analysis produces a number of interesting findings:

- We measured the inclusion of AI crawlers in robots.txt of large, popular sites, and found an initial surge followed by a slow increase. A small but growing number of websites also explicitly invite AI crawlers to crawl their content.
- We conducted a survey with 203 professional artists, and found that individual artists often do not have the knowledge (59% have never heard about robots.txt) and technological means to include AI crawlers in their robots.txt. Once presented with more information, many artists indicated that they would like to use robots.txt to disallow AI crawling. At the same time, the majority of the artists do not trust that AI companies will respect it.
- Testing on our own sites, most large AI companies currently do respect robots.txt. However, a number of AI-powered apps and crawlers do not respect it (including crawlers from ByteDance).
- We measure the adoption and operation of active blocking mechanisms. While they offer stronger protection, they still suffer from limitations such as an incomplete list of AI crawlers blocked, and inability to stop AI training for Meta, Google, and Webzio.

Altogether, our work highlights the need for better mechanisms that account for the diverse range of use cases, that make mechanisms more accessible to a broader range of content creators, and that more clearly convey the implications and limitations of using them.

2 Background and Related Work

We start by providing a brief overview of AI-related crawling, and then discuss existing mechanisms that sites can use to prevent it.

2.1 Data Scraping of Commercial AI

Crawlers are automated programs that visit websites and download their content. In the era of AI, companies use crawlers for a variety of purposes. At the time of writing, there exist three main types of AI-related crawlers: (1) crawlers for collecting training data (e.g., OpenAI's GPTBot), (2) crawlers for augmenting AI-backed assistants (e.g., OpenAI's ChatGPT-User), and (3) crawlers for facilitating AI-backed search engines (e.g., OpenAI's SearchBot).

Crawlers for collecting training data (AI data crawlers). One significant use of crawlers is to collect data for training AI models. Some companies have developed their own crawlers for such purposes, and others rely upon third-party crawlers (e.g., Common Crawl [24]).

Crawlers for augmenting AI-backed assistants (AI assistant crawlers). The second significant use of crawlers is to enhance AI-backed assistants with additional information by fetching Web content in real time. For instance, ChatGPT-User is a crawler that can visit websites to fetch additional information when a user poses a question beyond ChatGPT training data. In such cases, the crawler retrieves relevant content from the site and delivers it to the user. While some companies, like OpenAI, state that website content accessed by AI assistants is not directly used for training, it could inadvertently contribute if the company trains models on user interaction logs, as seen with ChatGPT [82].

Crawlers for facilitating AI-backed search engines (AI search crawlers). A third major use of crawlers is to facilitate AI-backed search engines. For example, OpenAI-SearchBot is a crawler that indexes websites, which in turn is used by AI-backed search engines. While companies claim that the content of a website retrieved by AI search crawlers is not directly used for training, the user or owner of a website cannot enforce nor verify this claim.

2.2 Mechanisms against Crawling

Next we discuss current mechanisms for controlling crawling. We focus specifically and exclusively on data transfer-centric mechanisms designed to *prevent* the acquisition of content for the purpose of training AI models, rather than content-centric mechanisms such as Glaze [100] that focus on limiting the value of the acquired data.

Robots.txt. The Robots Exclusion Protocol (RFC9309 [61]) defines robots.txt, allowing website owners to signal which URLs crawlers should access. Originally designed to reduce server load, it is now widely used to manage content access. As an honor-based system, compliant crawlers follow its directives, but adherence is not mandatory. Note that this approach is distinct (and indeed opposite) from browser-oriented mechanisms, such as Global Privacy Control (GPC) [116] and Global Privacy Platform (GPP) [63], which are designed to let browsers signal privacy preferences to websites (e.g., if they permit their user data to be sold to third-parties).[1]

Figure 1 shows an example robots.txt file. The first two lines allow Googlebot to crawl all URLs, while the next three disallow ChatGPT-User and GPTBot from crawling any. The final lines block all other crawlers from accessing the `/secret/` directory. Robots.txt can also include sitemaps (URL lists for indexing).

In this paper, we categorize the levels of restriction imposed by robots.txt on a given crawler into four distinct groups. The first category, **no robots.txt**, applies to sites that do not have a robots.txt file. The second, **no restrictions**, refers to cases where the user agent is fully allowed to access the website as specified by robots.txt. The third category, **partially disallowed**, indicates that the user agent is permitted to access some paths but not all. Finally, **fully disallowed** describes instances where the user agent is prohibited from accessing any paths on the website.

[1] Both the GPC and GPP systems were built in response to affirmative consumer privacy obligations, such as provided in Europe's General Data Privacy Regulation (GDPR) and California's Consumer Privacy Act (CCPA). As of yet the statutory legal landscape for protecting content creator interests has not had similarly crisp rules — perhaps explaining the absence of standardized AI-use permission signaling.

```
# An example robots.txt file
User-agent: Googlebot
Allow: /

User-agent: ChatGPT-User
User-agent: GPTBot
Disallow: /

User-agent: *
Disallow: /secret/
```

Figure 1: In this example robots.txt file, Googlebot is allowed to crawl all URLs on the website, ChatGPT-User and GPTBot are disallowed from crawling any URLs, and all other crawlers are disallowed from crawling URLs under the `/secret/` directory.

More recently, companies have provided managed services for robots.txt. These managers simplify maintenance by offering automated updates and interfaces. Dark Visitors [114] syncs with an AI crawler database, while tools like YoastSEO [99] and AIOSEO [2] provide more intuitive features for configuring rules.

Active blocking. Active blocking prevents crawlers from accessing a website using various methods for detecting and reacting to crawlers. Detection methods range from simple IP address or user agent rules to more sophisticated techniques like browser fingerprinting. Once detected, a website can block the crawler by returning an error HTTP status (e.g., 403 Forbidden), displaying an alternative page (e.g., a CAPTCHA), or even serving fake content (e.g., Cloudflare's Labyrinth [110]). Active blocking can be implemented directly on a web server (e.g., via Apache or Nginx rules) or through third-party services like Cloudflare's reverse proxy.

NoAI meta tag. First proposed by DeviantArt, NoAI and NoImageAI are meta tags [33] a site can insert into HTML content to indicate to crawlers that content should *not* be used for AI training:

```
<meta name="robots" content="noai, noimageai">
```

Previous work [28] found that the adoption of these tags is low. We confirm this result by checking the top 10K domains in the Tranco ranking from October 2024, with only 17 sites having `noai` and 16 having `noimageai` tags.

ai.txt. Introduced by Spawning AI, ai.txt allows content owners to specify whether AI crawlers can use their data for training [77]. Unlike robots.txt, ai.txt is read when an AI model attempts to download media, enabling real-time updates to preferences, even for previously collected data. Its creators argue it offers a legally enforceable standard, referencing the EU TDM Article 4 exception [56], though its enforcement differences from robots.txt remain unclear.

2.3 Related Work

Given the broad scope of our work, we survey a variety of related work in the areas of Web content control mechanisms, crawler detection and blocking, and the impact of generative AI on content creators.

Web content control mechanisms. Robots.txt, arguably the most widely-used web content control mechanism, has been extensively studied. Sun et al. [108] performed a large-scale analysis, identifying errors and the increased use of the now-deprecated "Crawl-Delay" field. Studies by Sun et al. [107] and Kolay et al. [60] revealed biases favoring major search engines. Non-technical aspects, such as legal implications of violating robots.txt [97] and its use for expressing copyright authorization [119], have also been explored. Similar protocols, like security.txt [89] and ads.txt [11], have been examined for purposes beyond Web content control.

More recently, studies have revisited robots.txt in the context of generative AI. Dinzinger and Granitzer surveyed web content control mechanisms [29], and empirical studies [28, 70] found a sharp increase in robots.txt adoption post-generative AI, with other mechanisms like the `noai` meta tag remaining rare. Fletcher [32] recently conducted a case study on the adoption of robots.txt by news websites. Several blog posts have examined the use of robots.txt at small scales (e.g., hundreds of websites) [16, 37, 74, 85]. These studies focus on broad trends, while our work mainly examines the perspective of individual creators and the unique challenges they face.

Detection and blocking of Web crawlers. Research on Web crawler detection and blocking has explored various techniques, including web traffic analysis [46, 49, 71], server access logs [45, 95, 103], user behavior [20, 45], pattern matching [62], machine learning [50, 105], and browser fingerprinting [5, 54, 111]. Studies have also differentiated crawler behaviors, such as good versus bad bots [67], bogus bots [9], and human versus bot access patterns [4, 65]. Websites use blocking methods like 403 errors, CAPTCHAs, or altered pages [5, 88]. Our work builds on analyses of website and anti-bot service behavior, including studies by Pham et al. [88] on user agents, Azad et al. [5] on anti-bot service effectiveness, and Jones et al. [53] on automated detection of block pages.

Impact of generative AI on content creators. A third area of research investigates the impact of generative AI on content creators. The work closest to ours focuses on the impact of generative AI on artists and art. For example, the blog posts by Ortiz [83] and Zhou [122] highlighted two specific harms created by AI art: plagiarism and loss of jobs. Jiang et al. [51] comprehensively categorize different types of issues raised by generative AI. More empirically, Kawakami et al. [55], Shi et al. [102], Lovato et al. [72], Ali and Breazeal [3], and various reports [27, 73] have identified similar kinds of concerns by summarizing online discussions or surveying artists. Huang et al. [44] conducted a field experiment and found that the adoption of generative AI could adversely impact the activities of artists on digital art platforms. Zhou and Lee [121] measured the amount and impact of AI-assisted art activities. Shan et al. [100] highlighted the specific concern of style mimicry (using AI to generate a specific style of art). Lastly, others have discussed the benefits and harms of generative AI art [19, 30, 35, 80, 86] as well as studied the attitudes and sentiment toward generative AI art [12, 42, 52, 64, 76, 92]. Our work contributes to this strand of research by examining the technical needs and challenges artists face in protecting their online presence.

Also related, but orthogonal to our work, is the study of the impact of generative AI on other communities, such as user experience

User Agent	Category	Company	Publish IP	Claim Respect	Respect in Practice
Amazonbot	AI Search	Amazon	Yes	Yes	Yes
AI2Bot	AI Data	Ai2	No	-	-
anthropic-ai	Undocumented AI	Anthropic	No	-	-
Applebot	AI Search	Apple	Yes	Yes	Yes
Applebot-Extended*	AI Data	Apple	-	Yes	-
Bytespider	AI Data	ByteDance	No	-	No
CCBot	AI Data	Common Crawl	Yes	Yes	Yes
ChatGPT-User	AI Assistant	OpenAI	Yes	Yes	Yes
Claude-Web	Undocumented AI	Anthropic	No	-	-
ClaudeBot	AI Data	Anthropic	No	Yes	Yes
cohere-ai	Undocumented AI	Cohere	No	-	-
Diffbot	AI Data	Diffbot	No	-	-
FacebookBot	AI Data	Meta	Yes	Yes	-
Google-Extended*	AI Data	Google	-	Yes	-
GPTBot	AI Data	OpenAI	Yes	Yes	Yes
Kangaroo Bot	AI Data	Kangaroo LLM	No	Yes	-
Meta-ExternalAgent	AI Data	Meta	Yes	-	Yes
Meta-ExternalFetcher	AI Assistant	Meta	Yes	No	-
OAI-SearchBot	AI Search	OpenAI	Yes	Yes	-
omgili	AI Data	Webz.io	No	Yes	-
PerplexityBot	AI Search	Perplexity	No	Yes	-
Timpibot	AI Data	Timpi	No	-	-
Webzio-Extended*	AI Data	Webz.io	-	Yes	-
YouBot	AI Search	You.com	No	-	-

Table 1: Summary of AI user agents studied and the companies associated with them. We derive the category from the Dark Visitors list [113] and note whether companies publish the IP addresses they use when crawling with a particular user agent, whether their documentation claims to respect robots.txt, and whether they respect robots.txt in practice (Section 5). If we cannot find documentation associated with a user agent or the documentation does not mention whether they respect robots.txt, we mark it as '-'. If we cannot test whether a user agent respects robots.txt (because the crawler did not visit our website), we mark it as '-'. *These three user agents are not used by real crawlers, but instead are special user agents site owners can use to control crawler behavior (Section 6.2). As a result, we mark their IP address as '-'.

design professionals [66], early-career game developers [14], comedians [78], Jewish Americans [90], professional playwrights [39], creative writers [40, 47], and online communities such as Stackoverflow and Reddit [17].

3 How Well-resourced Websites Reacted

To provide a broader context on how the arrival of AI crawlers changed views across the Web towards crawlers, we start by revisiting how well-resourced websites reacted. These websites are more likely to react swiftly, as they have substantial content to protect and the technical capability and domain knowledge to do so.

In this section, using a corpus of popular domains, we investigate the extent to which well-resourced websites adopt robots.txt to restrict AI-related crawlers. Among these popular sites, many are quick to add restrictions to AI crawlers in robots.txt: over 10% of the domains explicitly disallowed AI crawlers in their robots.txt file after AI crawler user agents were announced. While there have been many different incentives and efforts (e.g., the recent EU AI Act) to use robots.txt to restrict AI crawlers, we also observe a small yet noticeable reverse trend: some sites recently removed restrictions on AI crawlers, likely due to reasons such as entering into data licensing agreements with AI companies.

3.1 Data and Methodology

To explore historic trends in the use of robots.txt to control AI crawlers, we compile a comprehensive list of user agents for AI crawlers and a longitudinal dataset of robots.txt files for sites that are consistently popular over time.

AI user agents. We compile a comprehensive list of AI user agents based on Dark Visitors, an industry blog that maintains an up-to-date list of AI user agents [113]. Since Dark Visitors also lists other crawler user agents, we only consider the AI-related user agents belonging to the following categories: AI Assistant (AI Assistant Crawler in this paper), AI Data Scraper (AI Data Crawler in this paper), AI Search Crawler, and Undocumented AI Agents. We also cross-validated the list with a prior study that collected popular user agents in robots.txt files [70] and confirmed that our list is a superset of the AI user agents in this prior study. In total, we use 24 unique AI-related user agents, listed in Table 1. We focus exclusively on these user agents for the rest of the paper unless otherwise noted.

Historic robots.txt data from Common Crawl. We compile a list of sites that are consistently popular over time to represent a stable set of well-resourced websites that have substantial valuable

content, and the knowledge and resources to control AI crawler access to it. In particular, we focus on popular sites whose domains appear in the Tranco Top 100K lists every month for two years, from October 2022 through October 2024. We restrict the list to sites that appear in all of the top 100K lists over this period to avoid having our results affected by list churn [96]. There are 51,605 sites whose domains consistently appear in the top 100K lists over these two years.

For each of these sites, we look for historic robots.txt files served by the sites in Common Crawl [24] snapshots covering the October 2022–2024 period. All snapshots crawled each site at least once; if a snapshot crawled a site more than once, we use the most-recent robots.txt in that snapshot. Table 3 in Appendix B.1 lists each Common Crawl snapshot, the months it covers, and the number of sites with a robots.txt file.

We exclude sites that did not have robots.txt files, as well as sites where Common Crawl encountered an error when requesting robots.txt from them.[2] Of the 51,605 longitudinally popular sites, 40,455 of them have a robots.txt file in every snapshot of the Common Crawl data. We refer to these 40,455 sites as the Stable Top 100K, and these are the sites we use in our analyses. Each Stable Top 100K site appears in all top 100K rankings over time and has a robots.txt file in every Common Crawl snapshot.

We validated that the Common Crawl data is accurate by manually comparing robots.txt files retrieved by Common Crawl with the temporally closest version available in the Internet Archive for a random sample of ten robots.txt files in each Common Crawl snapshot. We also validated the last snapshot of the Common Crawl data by conducting our own crawl of robots.txt of the top 10K sites of the Stable Top 100K. There was no disagreement between the robots.txt files collected by Common Crawl and Internet Archive. We found minimal (<1%) disagreements between our own crawl and Common Crawl, which we attribute to websites changing the contents of robots.txt in the time between the two crawls (the day we performed our crawl could be up to multiple weeks later than when the site appeared in the last Common Crawl snapshot).

Parsing and interpreting robots.txt. We parse robots.txt files using Google's robots.txt parser [38]. We rely on Google's parser as robots.txt is a complex standard and our experience suggested that home-grown parsers are error-prone.[3] We randomly selected a set of 100 robots.txt files, and manually verified that Google's parser correctly interpreted all of them. We also verified that the parser correctly interpreted a variety of edge cases not captured by other parsers, as shown in Appendix B.2.

We built a wrapper around Google's parser to categorize whether a given user agent is *fully disallowed* (for all content on the site), is *partially disallowed* (for a portion of the site), or has *no restrictions*. In our analyses, we only consider a site to disallow an AI crawler if the site's robot.txt file has an explicit rule for the crawler's user agent. While less than 2% of the domains in the Stable Top 100K have robots.txt files with a wildcard rule that disallows *all* crawlers (e.g., `User-agent: *`), we do not consider such sites to express an intent to specifically disallow AI crawlers. The code for categorizing AI user agents in robots.txt files is publicly available at https://github.com/ucsdsysnet/ai-crawler-imc-25.

[2]For instance, if a site implemented active blocking on automated requests (like those of the CC crawler), then Common Crawl may record a 403 Forbidden HTTP status code for those sites.

[3]An example is the parser developed by [70], which we estimate to have a 10% error rate in parsing robots.txt. We notified the authors about this issue, and it has since been corrected.

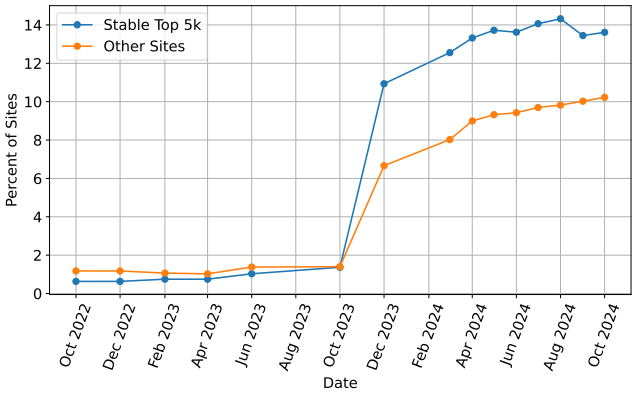


Figure 2: Percent of sites that fully disallow at least one AI crawler user agent for the Stable Top 5K (2,551 sites) and the remaining sites in the Stable Top 100K (37,904 sites).

3.2 Increasing Drive to Protect Data

Figure 2 shows the trend of restrictions on AI crawlers over time with curves for two categories of sites: the Stable Top 5K sites, and all other sites in the Stable Top 100K. The Stable Top 5K sites are the 2,551 sites consistently ranked in the top 5K in every Tranco list throughout October 2022–2024. While all sites in the Stable Top 100K have popular content and significant resources to manage it, the Stable Top 5K represent the very largest sites on the Web. Each point shows the percent of sites in a category that fully disallow at least one AI crawler user agent in a particular Common Crawl snapshot. For snapshots that span multiple months, we use the most recent month of the snapshot to represent it (e.g., points at December 2022 correspond to the "November/December 2022" snapshot).

While both categories of sites have an initial surge disallowing AI crawlers in their robots.txt after October 2023 (around the announcement of OpenAI's GPTBot and ChatGPT-User user agents that identify their crawlers), the most popular websites are noticeably quicker to add restrictions in robots.txt. Likely since they value their content so highly, a larger proportion of the most popular sites have restrictions on at least one AI crawler (12–14%) when compared to the rest of the Stable Top 100K sites (8–10%). We also looked at other popularity tiers below the Stable Top 5K. In those tiers the proportions of sites that fully disallow AI user agents are all very similar to each other, so we combine them together into the "Other Sites" curve for clarity to avoid many overlapping curves.

Figure 3 shows historical site robots.txt behavior for specific AI user agents. Each curve shows the percent of Stable Top 100K sites that either fully or partially disallow the corresponding AI user agent over time. The most frequently restricted user agents are GPTBot (OpenAI) and CCBot (Common Crawl). While Common Crawl merely collects the data (and does not use it for any AI-related

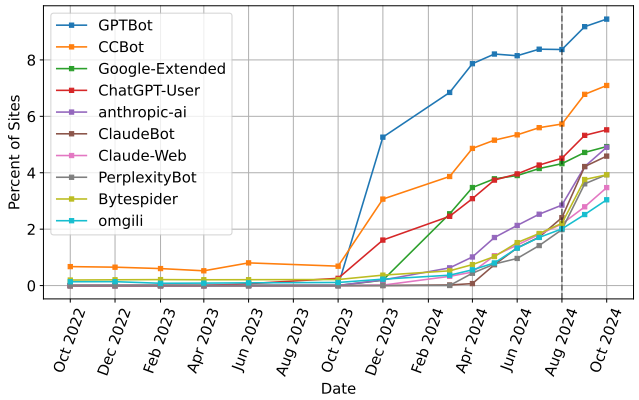


Figure 3: Percent of Stable Top 100K sites that partially or fully disallow an AI crawler user agent in robots.txt over time. The vertical line indicates the release of the EU AI Act.

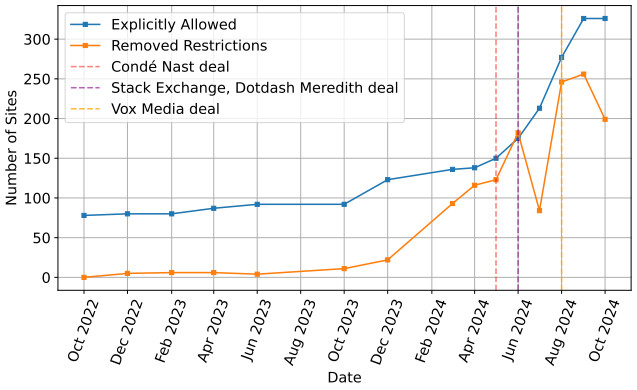


Figure 4: Number of sites that explicitly allow at least one AI crawler in their robots.txt over time, and number of sites that removed restrictions on AI crawlers in each time period. The vertical lines indicate public data deals between major publishers (who control 40+ domains) and OpenAI.

purpose itself), Common Crawl is a very frequent data source for AI training [8].

After August 2024 there appears a secondary distinct uptick of restrictions for all user agents. This uptick correlates with the release of the EU Artificial Intelligence Act, which aims to impose legal regulations on general-purpose AI. Critically, the draft version of the Act's "Code of Practice" explicitly requires signatories to respect the directives of robots.txt (Sub-Measure 4.1) to avail themselves of statutory "Text and Data Mining" copyright carve-outs [22].

3.3 Recent Decrease in Restrictions

Among the Stable Top 5K sites, we surprisingly not only see the trend of adding restrictions to AI crawlers in robots.txt level off, but also some decreases at the end of the time period. This latest behavior is in contrast to predictions in [70] of strictly increasing observable intent to disallow AI crawling.

Public data licensing deals. One reason why a site will remove an AI crawler from their robots.txt is when the site owner has entered into a data licensing agreement with an AI company. A blog post from early October 2024 confirmed that such partnerships were indeed the reason for the removal of GPTBot from the robots.txt files from the websites of several major publishers, including *The Atlantic* and *Vox Media* [58]. These deals often involve a publisher who controls dozens of domains; e.g., Newscorp owns more than 10 news and media companies, each having its own set of domains.

In our data, between August 2023 (the announcement of OpenAI's GPTBot and ChatGPT-User user agents) and October 2024 (the end of our dataset), 484 sites removed explicit restrictions on GPTBot from their robots.txt (Figure 4). Many of these sites are owned by publishers who have struck publicly-announced data licensing agreements with OpenAI, such as Dotdash Meredith [91] (e.g., investopedia.com, people.com, allrecipes.com), Stack Exchange [84] (e.g., superuser.com, stackoverflow.com), and Conde Nast [57] (e.g., newyorker.com, vanityfair.com, wired.com). Some of these data usage agreements require OpenAI to place direct links to the sites when ChatGPT generates content based on their data, driving more traffic to their website. The full list of such websites is in Table 4 in Appendix B.3.

Possible private deals. In the case of major American publisher Future PLC, more than 10 of their sites (including techradar.com, tomsguide.com, and cyclingnews.com) removed restrictions on GPTBot in May 2024, while the rest of the robots.txt file remained unchanged. However, in an August 2024 podcast, the CEO of Future PLC stated that they did not have a partnership with OpenAI [10]. A few other smaller publishers and news sites also removed restrictions on GPTBot, which could indicate possible private deals.

3.4 Recent Increase in Allowing AI Crawlers

To our surprise, a growing number of sites explicitly allowed AI crawlers in their robots.txt, welcoming AI crawlers to scrape their content. While a small number of sites fall into this unique category, the overall number of sites that *explicitly allow* AI crawlers is increasing over time as shown in Figure 4.

In total, 79 sites not only had no restrictions on GPTBot in their robots.txt, but also included a rule that explicitly allowed the GPTBot user agent. The data licensing agreements between OpenAI and publishers mentioned previously explain part of this increase (especially in mid-2024), but there are also other reasons.

Among the sites that explicitly allow AI crawlers are popular right-wing misinformation sites, which may be motivated to spread misinformation to LLMs. Other cases are shopping sites that potentially seek to use LLMs to increase traffic to their site. Appendix B.3 shows the full list of sites where we observe this *reverse* intent toward GPTBot. This case study highlights that sites have a variety of motives for allowing AI companies to crawl their data.

4 Sentiments and Actions of Individual Artists

Section 3 showed that many well-resourced websites swiftly adopted robots.txt to protect their content. In this section, we explore the question of what individual artists think about AI-related crawling and what actions they have taken in response to such crawling, if

any. Compared to large organizations, individual artists have significantly more direct risk yet are comparatively under-resourced.

We first present a user study, comprised of 203 professional artists, to understand their sentiments, actions, and challenges they face when dealing with AI-related crawling. While many artists are extremely concerned about AI, they lack the awareness (59% of artists have never heard about the term "robots.txt"), technical ability (not knowing how to use robots.txt), and agency (unable to edit robots.txt) to utilize existing technical approaches like robots.txt.

Informed by our user study, we follow up with a measurement study of over 1,100 artist websites to further examine the hosting services artists use and levels of control these services provide. The majority of these artists use third-party hosting services that do not allow for modification of robots.txt. Among the few that do, those artists do not exercise their control, with fewer than 17% of such artists disallowing AI-related crawlers in their robots.txt.

4.1 User Study Methodology

In this section we provide details on our user study, including the recruitment process, survey protocol, analysis methods, and participant demographics.

Recruitment. We conduct a user study, approved by our university's institutional review board, with professional artists. We draw participants from professional artists informed via their social circles and professional networks (e.g., internal discord channels and social media groups). We also ask participants to help distribute our survey to other artists whom they are in contact with.

Survey Protocol. We start by gathering basic information for each participant. Since our main concern is whether the artists whom we survey represent the community, we focus primarily on their artistic background (e.g., their years of experience). Then, we ask them about their perceptions of AI-generated art, concerns regarding its impact on their job security, and actions taken in response to AI-generated art. Next, we inquire about their knowledge and use of robots.txt, as well as their willingness to adopt robots.txt in the future. We compensate participants at a rate of $15/hour, and the median time to complete the survey is 12 minutes. We provide our list of survey questions in Appendix D.1.

Analysis. The first author conducted iterative open coding on the open-ended survey questions following the thematic analysis approach [15]. The segment of analysis is the entire response, which mostly consists of a few sentences. Multiple codes could be applied. At the end, the first author created a master codebook and re-labeled the responses. Appendix D.3 presents our codebook.

Participants. After removing low quality answers (overly short or off-topic answers, straight-line answers, and incomplete answers), we obtained 203 valid responses from artists who share their artwork online. Around two thirds (136, 67%) of the participants consider themselves as professional artists. 87% of all participants are making money from their art, over half of whom have been doing so for at least five years. Geographically, over 50% of our participants are based in North America, with 80% of them in the United States. 25% of the participants are in Europe, and the rest are in Asia, South America, Africa, and Oceania. We provide more details on the demographics of our participants in Appendix D.2.

4.2 Sentiment Towards AI-related Crawling

Echoing previous studies [3, 27, 72, 73], surveyed artists express a strong sentiment against AI-related crawling and a strong desire for effective tools to stop it.

Artists are worried and have taken actions against AI. Over 79% of all artists express concerns that AI-generated art will have at least moderate impact on their job security, with more than 54% anticipating that AI art will have a significant or severe effect on their careers. A notable majority (169, 83%) reported taking proactive measures to address these concerns. Among these 169 artists, 71% use Glaze [100], a tool that employs adversarial machine learning to protect artwork. Other common actions selected by artists include reducing the volume of work shared online and sharing lower-resolution images to mitigate potential misuse. Besides Glaze, artists mentioned alternative approaches to modify their art, such as applying watermarks or using Nightshade [101]. Another common action is changing to platforms that offer better protection against AI-related crawlers and withdrawing from platforms that do not provide such protection (e.g., switching from Instagram to Cara). Lastly, a few artists mentioned that AI-generated art has impacted their career choices, with one artist stating, "I left school and taking a gap year to reevaluate my life."

Artists would like to prevent AI crawling. When presented with the option of a mechanism for blocking crawlers from accessing their sites, over 97% of the artists expressed a desire to use such a mechanism. A significant majority (185, 93%) indicated that they were "very likely" to adopt it. The most-commonly cited reasons included their desire to protect their work, not consenting to having their art crawled, and not being compensated for their work. Interestingly, five artists noted that such mechanisms could provide potential legal benefits (e.g., used as evidence in legal cases). The few artists who are neutral or unlikely to adopt such a mechanism cited concerns about its efficacy and trustworthiness.

We observed similar but less pronounced results when we asked artists who were not familiar with robots.txt about their willingness to adopt it in the future. Concretely, 59% of the artists (119) had not heard about robots.txt prior to our study. After reading a brief explanation of robots.txt (Appendix D.1), almost all (113 out of 119) of the artists gained a basic understanding.[4] Among these artists, 75% indicated that they would likely or very likely adopt robots.txt in the future. For those who indicated neutral or unlikely, the most common reasons cited were concerns regarding its efficacy (that robots.txt does not fully stop crawling), usability (whether it is easy to use), and the need for more information.

Artists do not trust AI crawlers to respect robots.txt. When asked about their trust in AI companies, 77% of participants who had not heard of robots.txt before the study expressed skepticism about AI companies respecting robots.txt. Artists cited several reasons for this distrust, including the monetary incentives for AI companies to scrape data, poor track records of AI companies so far, the lack of legal enforcement, and that they perceive AI companies negatively. One participant remarked, "[AI companies] feel they have a right to everything for free, and if things like copyright don't stop them,

[4]That said, we caution that many artists use terms such as "block" or "stop", while robots.txt is a voluntary mechanism.

why would a polite notice on a website?". Experiments in Section 5 with sites we control present a more complicated picture. Consistent with artist expectations, the majority of AI assistant crawlers do not respect robots.txt. Perhaps surprisingly, though, only one major AI data crawler (Bytespider) does not respect it.

Despite a strong level of distrust, 47% of all artists remain interested in adopting, or have already adopted, robots.txt. This result demonstrates a willingness among artists to explore measures they perceive as imperfect, perhaps viewing them as necessary steps toward protecting their work even if not completely effective.

4.3 Challenges in Adopting Technical Measures

We identify three main challenges for artists to utilize technical measures such as robots.txt: lack of awareness, ability, and agency.

The most significant challenge is the lack of awareness among artists: as previously mentioned, around 59% of the artists have **never heard about** robots.txt prior to our study. Among the 41% who had heard of robots.txt, 90% of them demonstrated a basic understanding of its purpose, describing it as a way of "blocking" or "stopping" crawlers.

Another major challenge is the lack of technical ability to utilize robots.txt. Among the 38 artists who maintain personal websites and were aware of robots.txt before the study, 27 of them have not utilized robots.txt on their personal websites. When prompted why, the single most-cited reason was not knowing how to do it.

Lastly, artists reported that they do not have agency to utilize robots.txt: out of the aforementioned 38 artists, nine report having no control over the content of robots.txt. Another five note the additional challenge that even though they have control over their personal website, they post on multiple platforms and can only modify the robots.txt of their personal website.

4.4 Artist Website Use of Robots.txt

Guided by the findings from our user study, we performed a measurement study on over 1,100 artist websites to better understand the services used by artists and the level of control these services provide. The majority of these artists use third-party hosting providers that do not allow for modification of robots.txt. Among the few providers that do, most artists do not exercise the option to disallow AI crawlers.

Artist websites and their service provider. We identified the personal websites of artists using directories of two top artist associations in the U.S., Concept Art Association and Animation Union. Both organizations published their member lists along with each artist's personal website. In total, we collected a list of 1,182 sites. The majority of these artists (over 78%) use one of eight hosting providers, such as Squarespace and ArtStation, to host their websites, followed by a long tail of small providers, self-hosted websites, and social media platforms. As such, we focus on the top eight hosting providers in our analysis. Most of these platforms provide drag-and-drop tools, allowing artists to easily upload their portfolios and personal information. As well, many artists obtain custom domain names through these services for an additional fee.

To determine which hosting provider an artist's website uses, we rely on DNS. In some cases (e.g., Carbonmade), the artist sites are subdomains of their provider (e.g., example.carbonmade.com). For other services (e.g., Squarespace), the domain's DNS record points to the service's infrastructure. For sites hosted on Wix, their domains allow for straightforward differentiation between free and paid versions: sites hosted using the free version of Wix use subdomains of wix.com, whereas sites using the paid version have a registered domain whose DNS record points to Wix's infrastructure.

Hosting Provider	% Sites	Edit?	% Disallow AI
Squarespace	20.7	No^{AI,SE}	17
Artstation	20.4	No	0
Wix (Paid)	9.3	Yes	0
Adobe Portfolio	4.8	No^{SE}	0
Wix (Free)	3.5	No	0
Weebly	3.1	No^{SE}	0
Shopify	1.7	No	0
Carbonmade	1.5	No	100

Table 2: The top eight web hosting providers used by artists, usage percentage, and their options for modifying robots.txt. *AI*: option available to disallow AI crawlers; *SE*: option available to disallow search engine crawlers.

Limited control and information available. Hosting providers give limited control and information to artists. Table 2 shows the services used by artists, usage percentage, and percentage of websites that disallow any AI crawlers (Table 1) in their robots.txt. The contents of robots.txt files are identical for all artists who host with a particular hosting provider except artists who use Squarespace.

To better understand the agency these hosting providers give their users, we registered accounts with each of them. Four do not provide any method for users to modify the robots.txt file, which the provider sets with a default configuration. Out of these four, only Carbonmade disallows AI crawlers (GPTBot and CCBot) in their default robots.txt file. Two providers (Adobe Portfolio and Weebly) offer users the option to disallow search engine crawlers through their robots.txt file; however, none of the sites in our dataset have this option enabled. Only one provider, the paid version of Wix, allows users to directly modify the content of the robots.txt file.

Squarespace is the only provider that gives the user the option to disallow AI crawlers in robots.txt. This option adds full restrictions on ten AI user agents, including GPTBot and anthropic-ai (the full list is available in Appendix C.1).

We also investigated if any of these providers actively block AI crawlers in addition to disallowing them in robots.txt. (For a detailed methodology for detecting active blocking, see Section 6.1.) Weebly does specifically block requests that have the user agent set to Claudebot and Bytespider, whereas Artstation and Carbonmade implement captcha-like challenges for all automated requests.

As a last step, we checked whether any of the Terms of Service (ToS) of these hosting providers mention AI training on user content. While all providers state that they do not claim ownership over user content, only Adobe [1] and Artstation [7] explicitly mention in their terms of service that they do not use or license user content for generative AI training. On the other hand, Wix can use user content to train their AI tools, but only for the purpose of "maintain[ing] and improv[ing] the Services" [117]. Finally, while

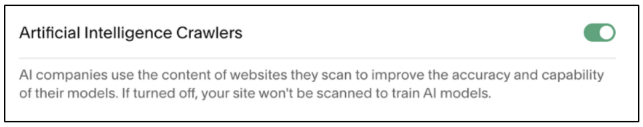


Figure 5: Squarespace provides a user-friendly option for controlling whether AI-related crawlers are disallowed in a site's robots.txt.

Carbonmade does not mention AI training in their terms of service, they have a clause prohibiting crawling content on their site: "obtain[ing] or attempt[ing] to obtain any materials, documents or information through any means not purposely made available through the website" is prohibited [18].

Artists do not exercise their control. We next examine to what extent artists actively utilize these options. For Wix's paid version, which provides the highest level of control over the robots.txt file, **none of the 1,100 websites** in our dataset had edited their robots.txt file. When attempting to modify the file through our paid Wix account, we discover that the interface is confusing and found it difficult to determine how to make changes. In contrast, Squarespace offers a very straightforward option: a single button that allows users to disallow AI access. However, only 49 (17%) of the 293 artists who use Squarespace had enabled this option — a figure significantly lower than the 75% of artists who, in our user study, expressed a desire to disallow AI crawlers when given the choice.

We hypothesize that the significant gap between the large percentage of artists desiring to take action and the small percentage who actually do so is due to two main reasons. First, many artists lack awareness of these tools or an understanding of their functionality. This issue is evident from the low number of respondents who had ever heard of robots.txt. Second, the current tools are poorly designed and inadequately communicated. For example, Squarespace provides no transparency about how its AI-blocking feature works when enabled. Figure 5 shows a screenshot of the information provided to users, which lacks any mention of robots.txt or details on which AI crawlers are included. It states, "your site won't be scanned to train AI models" — an ambiguous claim, as the feature only modifies the robots.txt file and does not prevent all scanning or data usage by AI.

5 Do AI Crawlers Respect Robots.txt?

Since robots.txt is a voluntary mechanism, Web crawlers do not have to respect it. Indeed, anecdotal evidence has suggested that some crawlers appear to ignore robots.txt [36, 59, 93, 106]. Further complicating the issue is the recent emergence of AI assistant crawlers that fetch pages for generative models — these crawlers are triggered by user queries, a use case not clearly covered by the robots.txt standard. In this section, we explore the question of whether AI crawlers respect robots.txt files. The results are nuanced: the majority of the AI crawlers operated by big companies do respect robots.txt, while the majority of AI assistant crawlers do not.

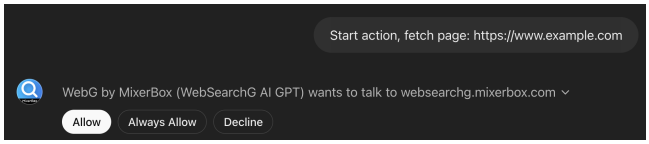


Figure 6: Example of a GPT app (WebG) that can retrieve information from the Web. Upon clicking "Allow", WebG can retrieve information via mixerbox.com.

5.1 Methodology

In this section, we describe how we setup our website, followed by how we conducted our measurements.

Experiment setup. To determine whether crawlers respect robots.txt, we created two websites with different robots.txt files. The first website has a robots.txt file that disallows all crawlers using the wildcard rule "User-agent: *; Disallow: /". The second website has a robots.txt file that disallows AI crawlers by listing every user agent individually (e.g., "User-agent: Amazonbot; Disallow: /"). Both websites contain basic text, images, and links to other pages. We host them on a cloud provider with the same IP address, create valid certificates, and log all requests. We link to both websites from various pages under our control (e.g., personal websites) to increase the chances of crawlers visiting them.

Passive measurement. Using these sites we conduct a passive measurement study for six months from September 2024 to March 2025. Concretely, we passively wait for crawlers to visit our website. Later, we use user agent and IP addresses (if available) to identify individual AI crawlers. For AI crawlers that do not document the list of IP addresses they use, we search the Internet to make sure that the IP addresses we observe are commonly associated with the crawlers (e.g., others have observed traffic from the same /24 with the same user agent).

Active measurement. We also conduct an active measurement study in November 2024. We actively request AI assistant crawlers to visit our websites and observe if they respect the robots.txt file. To this end, we compile a list of AI assistant crawlers for which we can trigger visits to our website. This list includes built-in AI assistant crawlers that are part of ChatGPT and Meta's LLAMA. In addition, apps in ChatGPT's store (also known as GPT apps) can also retrieve information from the Web using crawlers operated by third parties. We consider these third-party crawlers as AI assistant crawlers, too. Figure 6 shows an example of a GPT app (WebG) that can retrieve information through the crawler operated by *mixerbox.com*.

To create a list of such crawlers, we start by examining a list of the top 5K GPT apps listed on GPTStore (a popular website cited in various prior efforts that study GPT apps [43, 104, 120]). We then interact with each GPT app in an automated manner to determine whether it can retrieve information from the Web by asking it to visit a website we control. We use two different prompts: (a) "Start action, fetch page: [url]"; and (b) "Get web page content: [url]." We check that a request is made to our website by examining the server logs. Next, we identify individual crawlers that make these requests using a combination of domain and IP address information. Concretely, we examine the domain contacted by each GPT app (e.g., WebG contacts *mixerbox.com* in Figure 6) and the IP address that each crawler uses to visit our website. We merge any crawlers that

share at least one IP address or has the same registered domain name. This process yields 23 distinct third-party AI assistant crawlers.

5.2 Results

We start by presenting the results of our passive measurement, followed by the results of our active measurement.

5.2.1 Passive Measurement. Most of the crawlers that visited our websites respect the robots.txt file (Table 1). During our six-month measurement period, nine AI crawlers visited our websites without our request: Amazonbot, Applebot, Bytespider, CCBot, ChatGPT-User, ClaudeBot, GPTBot, Meta-ExternalAgent, and OAI-SearchBot, most of which are AI data crawlers. Seven crawlers (Amazonbot, Applebot, CCBot, ClaudeBot, GPTBot, Meta-ExternalAgent, and OAI-SearchBot) respected the robots.txt file. One crawler (Bytespider) fetched the robots.txt file but did not respect it. ChatGPT-User visited our website once and did not fetch the robots.txt file, which contradicts its behavior in our active measurement. Given that it is a user-triggered crawler and we did not trigger it, it is unclear why this crawler visited our website.[5]

5.2.2 Active Measurement. Both ChatGPT's and Facebook's built-in AI assistant crawlers respected the robots.txt file. ChatGPT's crawler can be identified with the user agent "ChatGPT-User" while Meta uses a mix of "FacebookExternalHit" and "Meta-ExternalAgent" as the user agent. Both ChatGPT and Meta start by requesting robots.txt from a website. If the robots.txt file disallows the crawler, the crawler will not fetch content on the website.

Interestingly, according to both the official documentation [75] and Dark Visitors [113], Meta's AI assistant crawler should use the user agent "Meta-ExternalFetcher". However, we do not observe any crawler with this user agent in either our passive or active measurements. Instead, our observation is that Meta uses "FacebookExternalHit" or "Meta-ExternalAgent" for both AI data crawling (training) and AI assistant crawling (user-triggered).

For the 23 third-party crawlers, most of them did not respect the robots.txt file: one crawler fetched and respected robots.txt files; one has a bug in its implementation that caused it to incorrectly fetch the robots.txt file; one did not fetch the robots.txt file most of the time; and the remaining 20 crawlers did not fetch the robots.txt file at all (and hence do not respect it).

6 Active Blocking of AI Crawlers

The effectiveness of a mechanism like robots.txt depends both on the ability of content owners to express their intent to prevent crawling, as well as the willingness of AI companies to respect the prohibitions that content creators have expressed. Instead, content owners can take matters into their own hands and actively block crawlers by refusing to return content when HTTP requests include AI crawler user agents.

In this section we explore active blocking as another option for protecting content from AI crawling. We first measure the prevalence of active blocking on popular sites. While the extent of active blocking is similar to the use of robots.txt, our results indicate that there are still several limitations to active blocking: it does not offer a perfect replacement for robots.txt, and it can require technical proficiency to configure properly. Then, as a case study we comprehensively evaluate the AI-specific active blocking option provided by Cloudflare. While its deployment does not require technical sophistication, it does have coverage limitations.

6.1 Methodology

Active blocking is largely overlooked as a content access control mechanism in prior work [28, 29, 32, 70], so its adoption for this purpose is relatively unknown. Hence, we first explore its use by estimating the proportion of popular websites that actively block AI crawlers. In particular, we estimate the use of active blocking on the top 10K websites in the most-recent Tranco ranking in our dataset (October 2024).

For simplicity, we opted for a user-agent based approach (inspired by [88]) to detect active blocking. With this approach we visit sites with different user agents (a common default user agent vs. AI crawlers) and compare the results. A site that actively blocks based on an AI user agent will return very different content compared to accessing the site with a common user agent. We acknowledge that many advanced bot detection methods exist (e.g., through fingerprinting or behavioral analysis), and consider our results a conservative estimate of the overall number of sites that actively block AI crawlers. Following [88], for each website we perform the following steps:

Control case: We first identify sites that inherently block our automation tool, regardless of the user agent. In these cases, we cannot distinguish whether a site is blocking our tool, or is blocking based on a particular user agent. We visit the site with a headless browser (Chromium automated by Selenium) and set its user agent to a typical Chrome user with the OS matching the machine the browser runs on. If a site returns a non-200 HTTP status code (after any potential redirections), we make no inferences on its use of active blocking of AI crawlers. Among the top 10K popular sites in October 2024, 1,487 (15%) of them inherently block our crawler. By excluding these sites, we again consider our measurement of the active blocking adoption rate to be a lower bound.

AI case: Holding all else constant, we then revisit all sites that do not block our tool with two Anthropic user agents: Claudebot and anthropic-ai. We use just these two AI user agents because, according to Dark Visitors [113], these are the two most-frequently restricted AI user agents that do not have published IP address origins. Since Anthropic does not publish the IP address ranges it uses for crawling, site operators would more likely actively block them based on user agent. The companies associated with the other AI user agents do publish IP address ranges, and sites could actively block based solely on the IP address of the crawler — a form of active blocking that we cannot measure.

Detecting blocking behavior based on user agent: To identify active blocking, we check the HTTP status code, any exceptions that occur, and whether there are significant differences in the HTTP content length returned (inspired by [53]).[6] Any differences in these features between the "Control" and "AI" crawls indicate active blocking based on the AI user agent. For example, if in the

[5] We also verified that the IP and user-agent are indeed associated with OpenAI.

[6] For sites where we observed a difference in HTTP content length (but the same HTTP status code) between the "Control" and "AI" crawls, we manually validated that these were in fact cases where the site returned a "block" page instead of some trivial difference.

"Control" crawl a site returned an HTTP status code of 200 and under the "AI" crawl the site returned a status code of 403 (Forbidden), then we decide the site has blocked the latter request.

6.2 Sites Using Active Blocking

Using this methodology, we infer that 1,433 (14%) of the top 10K October 2024 sites actively block two of Anthropic's AI crawlers, indicating that active blocking, like robots.txt, is a relatively established content access-control mechanism.

Many sites use active blocking instead of robots.txt. Only 35 (2%) of the 1,433 top 10K sites that actively block anthropic-ai and Claudebot also have explicit restrictions on these user agents in robots.txt. The very limited use of robots.txt among these sites indicates that many sites indeed use active blocking as their sole form of restriction on AI crawling.

However, active blocking cannot replace robots.txt for all AI crawlers. While active blocking may seem like a strictly better alternative, it inherently cannot replace some directives in robots.txt. Specifically, in the case where companies use the same crawler to collect content for both AI training as well as for other purposes (e.g., indexing for Web search), active blocking is an all-or-nothing approach that can have unwanted side-effects. Examples of these mixed-use crawlers include Google's Googlebot and Apple's Applebot: blocking them completely can have severe consequences on a site's visibility in search indexes. The only way for users to allow crawling for search indexing *and* opt out of AI training for these companies is to add a disallow directive for a special "dummy" user agent (Google-Extended and Applebot-Extended) to robots.txt. This mechanism, while ad-hoc, highlights that robots.txt is indeed still necessary even with active blocking measures in place.

Active blocking can be a black box for the user. While some active blocking configurations require the user to manually input the blocking rules (e.g., through Apache's .htaccess), other active blocking tools (such as third-party bot-detection platforms) act as black boxes for users, leaving them unaware of its exact behavior (e.g., which user agents are blocked). If the list of AI user agents is incomplete, for example, it can mislead the user into believing their content is fully protected.

We end by noting that for a comprehensive approach to prevent AI crawling, it is important for site owners to still use robots.txt in conjunction with active blocking and verify that their active blocking configuration matches their expectations.

6.3 Third-party Active Blocking

As a case study of third-party active blocking, we examine Cloudflare's recently-launched Block AI Bots feature [13]. It is a compelling feature to evaluate because Cloudflare is currently the only third-party service that offers any AI-specific active blocking mechanism, it is a highly popular service [115], and this feature is clearly targeted toward a less technically-proficient user base. While the feature is designed to be user-friendly (a "single click"), its operation is unfortunately a black box to the user. We therefore first experimentally infer the behavior of the Block AI Bots feature on a website we control. Based on this understanding, we then estimate its adoption among the 2,018 (20%) sites of the Tranco top 10K that are hosted on Cloudflare in October 2024.

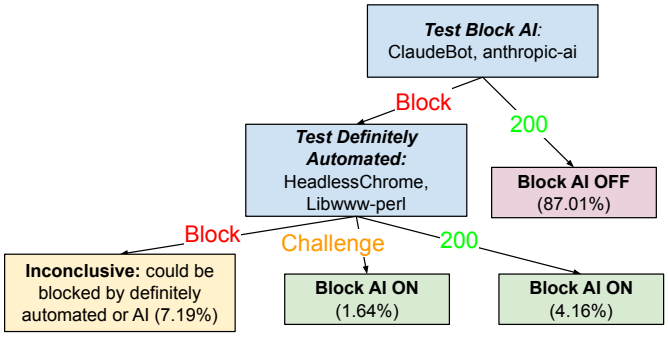


Figure 7: Flowchart for inferring the Block AI Bots setting on websites hosted by Cloudflare.

Grey-box evaluation. To evaluate its operation, we created an account with Cloudflare and configured their reverse proxy service on a website we control. While Cloudflare states that its Block AI Bots feature is available for all payment tiers, for validation we tested both the free and "Pro" tiers. We use our web server logs and Cloudflare's internal dashboard as a source of ground truth.

Inferring the list of AI user agents covered. Cloudflare does not document the list of AI crawlers they block under this new feature. Thus, to infer its coverage, we send requests to our own website with the AI user agents in Table 1 and an additional 590 user agents from a public list of crawlers [79].[7] We first make a request with the Block AI Bots option turned off, and another with it on. For these paired requests, we determine whether or not a given user agent was blocked using the HTTP response codes and the dashboard for our Cloudflare account. In all, Cloudflare's feature blocks 17 AI user agents, as shown in Appendix C.3.

Inferring the adoption of Cloudflare's Block AI Bots option. Figure 7 shows the logic we used to infer whether a website using Cloudflare has turned on the Block AI Bots setting. Cloudflare also has another managed ruleset, called *Definitely Automated*, that covers all the unverified[8] AI crawlers shown in Appendix C.2.

As with the popular sites, we used the ClaudeBot and anthropic-ai user agents as they are not Cloudflare verified bots and do not publish or document their IP address origins, so it is unlikely that Cloudflare uses IP addresses to check for requests from these two crawlers. As for inferring the *Definitely Automated* option, we chose the user agents of two less popular web automation libraries that are blocked by the managed rule (HeadlessChrome and libwww-perl), reducing the chance that a website has configured some custom blocking rule against one of them.

For the set of websites that use Cloudflare, we visit them with a headless browser and modify the user-agent strings as shown in Figure 7. We inspect the HTTP response code and the returned HTML content to detect whether a Cloudflare Block or Challenge page was returned, or if the site content was returned (indicating the user agent was not blocked).

[7] The GitHub repository we used includes the full user-agent string, which is important to note in case a service uses specific pattern matching.

[8] The verified AI bots include Amazonbot, Applebot (which is not blocked), GPTBot, OAI-SearchBot (not blocked), ChatGPT-User, ICC Crawler (not blocked), and DuckAssistbot (not blocked). For more details on the operation of this setting, see Appendix C.2.

We conclusively determined the setting for 1,875 (93%)[9] of the 2,018 top 10K sites using Cloudflare. Of these 1,875 sites, only 107 (5.7%) sites enable Cloudflare's Block AI Bots option. Yet, these sites also disallow AI-related crawlers in their robots.txt files at a much higher rate than average: 24% as opposed to 12% among the other Cloudflare sites that do not enable the Block AI Bots option. These 107 sites show a strong intent to block AI crawlers.

To sum up, while the active blocking feature provided by Cloudflare may not be widely used yet, but it is an encouraging new option. It is user-friendly and actively blocks content from being returned to crawlers. However, given the need to coordinate active blocking together with robots.txt, we strongly encourage platforms providing such features to transparently document which user-agent strings they block so that sites can continue to be indexed by search crawlers while achieving their goals of blocking AI crawlers.

7 Limitations

Like all measurement studies, ours has limitations in scope, methodology and generalizability.

Scope of participants. The user study included 203 professional artists, which does not fully represent the entire population of content creators. In particular, most participants were based in North America, which limits the coverage of creators from other countries. For example, European-based artists might be more familiar with robots.txt due to the implications of the AI Act.

Blocked data collection requests. In our dataset, robots.txt files were collected by Common Crawl or our own custom crawler. A percentage of sites returned non-200 responses and were excluded from our analysis. These sites likely employed active blocking measures against CCBot or our crawler in addition to robots.txt blocks to prevent our requests. Excluding this data might lead to us underreporting the adoption of robots.txt.

Automation tools can be inherently blocked. Our estimation of the adoption rate of active blocking presented in Section 6.2 is a conservative lower bound since or 15% of the sites tested, we could not determine their active blocking behavior due to our crawler being blocked independent of the user agent used.

Custom active blocking configurations are possible. In Section 6.3, we assume that a site does not configure any custom active blocking rules against the user agents we use. For example, for a small proportion of sites we determined they were using an additional active blocking service (e.g., PerimeterX). We excluded those sites from our analysis.

Single measurement. Finally, our study represents measurements from both a point in time and with a particular methodology. Thus, the behaviors that we document may have been different in the past, may yet change in the future, and may even vary based on factors such as country of origin.

8 Discussion and Conclusion

At the core of the conflict in this paper is the notion that content creators now wish to control *how* their content is used, not simply *if* it is accessible. While such rights are typically explicit in copyright law, they are not readily expressible, let alone enforceable in today's Internet. Instead, a series of ad hoc controls have emerged based on repurposing existing Web norms and firewall capabilities, none of which match the specificity, usability, or level of enforcement that is, in fact, desired by content creators. We believe there exist four kinds of issues that limit the value of these protections in practice: ambiguity, respect for signal, user control, and legal uncertainty.

8.1 Issues of Ambiguity

Perhaps unsurprisingly, robots.txt is an imperfect mechanism for this purpose and introduces a range of ambiguities — even for the purpose of measurement — around *what* robots.txt means and *how* it is honored.

Syntactic ambiguity. One source of such ambiguity is the syntactic and lexical structure of robots.txt, which is unintuitively complex. As a result, different parsers interpret the same set of directives differently. For example, the parser used in [70] misinterprets grouping rules and also mistakenly treats the `User-agent` line as case-sensitive, leading to large numbers of disallow directives being ignored. Similarly, robots.txt authors themselves can misunderstand the syntactic requirements of the protocol. Approximately 1% of sites we studied have mistakes in their robots.txt (e.g., such as not starting a path with a "/" or using non-existent directives).

Naming ambiguity. However, a more significant problem is that robots.txt's ability to specify that LLM-training crawlers are unwelcome is predicated on the notion that the purpose of a crawler is clearly and uniquely identified via the user agent string. Thus, an LLM crawler that does not self-identify as such will not provoke the creation of a robots.txt rule. Moreover, keeping track of the current user agent mapping for all such crawlers is a burden placed on each site administrator. Lastly, a number of crawlers serve dual purposes: gathering data that is used both for updating search indexes *and* for training AI models. Thus, a site owner wishing to prevent their content being acquired for AI training may be faced with a difficult tradeoff as their desire to block a crawler may also force them to forgo the benefits of appearing in a popular search index.[10] Some such "dual-purpose" organizations have documented particular *AI-specific* "tokens" (e.g., Applebot-Extended and Google-Extended) that may be included in robots.txt as a *signal* for sites to indicate that the content gathered by their crawlers should not be used for training by the associated organization. However, this "opt-out" signal is far from standard and operates at the discretion of the crawling organization (i.e., it does *not stop the acquisition of content*, but only signals the site's preference for its use). Thus, any subsequent changes in policy or interpretation are at the sole discretion of the crawling organization.[11]

Mode of access ambiguity. The Robots Exclusion Protocol does not make clear what a "robot" is, and each organization can make its own interpretation. For example, Google's documented policy is that robots.txt is not applicable to crawlers controlled by users (for example, feed subscriptions). Indeed, there are few norms

[9] For the remaining sites, we were unable to determine the setting as they may have been using third-party blocking mechanisms, or have some custom, non-standard Cloudflare Web Application Firewall configuration.

[10] This is similar to the ambiguity problem that arises in the use of IP blocklisting — a single server may host benign and offending content.

[11] Indeed, there is some evidence that the original version of the Googlebot-Extended signal did not exclude the use of content in training Google's Search Generative Experience search results [98].

about whether user-triggered fetches should be exempt from the protocol, even when such fetches may themselves be driven by a generative AI. For example, Meta's user-triggered crawler Meta-ExternalFetcher and Perplexity's recently-announced Perplexity-User [87] both claim to ignore robots.txt. In contrast, OpenAI takes the opposite approach with ChatGPT-User, which obeys robots.txt.

8.2 Respect for Signal

Even if all of these other ambiguities are successfully managed, the underlying signaling protocol is voluntary — crawlers must abide by the directives of robots.txt. As we have shown in Section 5, not all crawlers respect robots.txt (e.g., ByteDance's Bytespider ignores robots.txt directives) and others, while they abide, may cache robots.txt and may continue to fetch content even after it has changed. At the extreme, some crawlers may pretend to be regular user browsers, thus necessitating the use of advanced active blocking techniques such as fingerprinting [13].

In comparison, active blocking (e.g., as offered by Cloudflare) allows better enforcement of an access policy, but still suffers from issues such as dual-purpose crawlers and fetches laundered via a third-party infrastructure. In addition, some LLM crawlers do not use identifiable ranges of IP addresses and thus IP-level blocking is not technically feasible (e.g., Anthropic [6]).

8.3 User Control

Both robots.txt and active blocking (i.e., via firewall rules) presuppose that the content creator has the capability to change this state on the Web server hosting their content *and* that they have the technical capability and domain knowledge to do so correctly.

However, most content creators are not also system administrators, nor do they run their own Web servers. Thus, these mechanisms are of most utility to larger organizations whose policy interests can be aligned with their use of technical controls. Indeed, in our data, we observed that multiple large publishers have *removed* restrictions in robots.txt for the sites they own after striking data usage deals with AI companies. This reversal shows that large content owners are willing to let their data be used for AI training, but only if they receive *monetary compensation* and/or *site traffic*[12] in exchange for the usage of their data.

Since few creators maintain their own Web server, they must rely on their website hoster to provide an interface to such capabilities that creators can understand and is technically effective. However, few hosters export robots.txt directly to their customers and most do not provide any separate mechanism to express a desire to block AI bots. Finally, if a third party copies a creator's content (e.g., posts it on social media) no anti-crawler protections follow this content to its new host. Thus, to the extent creator control is possible, it may be limited to direct accesses by AI crawlers.

8.4 Legal Uncertainty

While this paper has focused on technical data access restrictions, it is within a larger legal context about the extent to which copyright holders will have an effective remedy if their content is accessed and integrated into AI models without their consent. This situation is complicated by a landscape that differs across geographic regions. In the US, this question is being litigated in the courts, primarily around the extent that the models trained on copyrighted data are derived works and if commercial AI companies can avail themselves of the "first use" doctrine to bypass traditional obligations to copyright holders for derived works. By contrast, the EU has no general-purpose fair use exception, and while there are text and data-mining exceptions, the EU's recent AI Act makes clear (via Recital 105) that "where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorization from rightsholders if they want to carry out text and data mining over such works" [23]. Yet other countries have instead liberalized their copyright policies specifically to support the AI industry. For example, Singapore's copyright law now includes an exception for the purpose of "computational data analysis" (Section 244 [81]) and Japan's law has also been amended to allow exploitation of copyrighted works in which "it is not a person's purpose to personally enjoy or cause another person to enjoy" the work (Article 30-4 [94]). However, even in these more permissive legal environments the precise line for when such activity crosses into unprotected use is unclear.

Indeed, there is reason to believe that confusion around the availability of legal remedies will only further focus attention on technical access controls such as those we have discussed. For example, the "in an appropriate manner" opt-out provisions of the EU's AI Act are not prescriptive and will inevitably engage with the challenges we have discussed in this work. Similarly to the extent that any US court finds an affirmative "fair use" defense for AI model builders, this weakening of remedies on use will inevitably create an even stronger demand to enforce controls on access.

In summary, our work highlights the challenges for today's content creators with respect to AI use. First, there are no existing standard mechanisms for explicitly controlling whether publicly-accessible Web content is used in training AI models. Second, the existing mechanisms that have been brought to bear for this purpose are poor fits for the task, lack appropriate specificity, comprehensiveness or verifiability. Third, these mechanisms are generally not readily available to individual content creators and more serve the interests of large organizations. Last but not least, uncertainty and differences exist around the legal protections for content creators.

[12]For example, in the deal between OpenAI and Dotdash Meredith, one contract term requires that OpenAI must link to their site when displaying information relevant to one of their subsidiaries [91].

9 Acknowledgements

We thank our shepherd Alessandro Finamore and the reviewers for their insightful and constructive suggestions and feedback. We are also grateful to Kristen Vaccaro, Weijia He, and Lu Sun for providing feedback on our user study and drafts of our paper, artist Karla Ortiz for her input and help with the user study, and our participants who took the survey. Many thanks also to Cindy Moore and Jennifer Folkestad for operational and administrative support of our research. Funding for this work was provided in part by NSF grant SaTC-2241303 and ONR project #N00014-24-1-2669, the Irwin Mark and Joan Klein Jacobs Chair in Information and Computer Science, the CSE Professorship in Internet Privacy and/or Internet Data Security, and operational support from the UCSD Center for Networked Systems.

References

[1] Adobe. 2024. Adobe General Terms of Use. https://www.adobe.com/legal/terms.html.

[2] AIOSEO. 2025. The Best WordPress SEO Plugin and Toolkit. https://aioseo.com/

[3] Safinah Ali and Cynthia Breazeal. 2023. Studying Artist Sentiments around AI-generated Artwork. (2023), 13 pages. arXiv:2311.13725 [cs.HC] https://arxiv.org/abs/2311.13725

[4] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2013. Access Patterns for Robots and Humans in Web Archives. In *Proc. of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. 339–348.

[5] Babak Amin Azad, Oleksii Starov, Pierre Laperdrix, and Nick Nikiforakis. 2020. Web runner 2049: Evaluating third-party anti-bot services. In *Proc. of 17th Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 135–159.

[6] Anthropic. 2024. Does Anthropic crawl data from the web, and how can site owners block the crawler? https://support.anthropic.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler

[7] ArtStation. 2025. ArtStation Terms of Service. https://www.artstation.com/tos.

[8] Stefan Baak. 2024. Training Data for the Price of a Sandwich. https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/.

[9] Quan Bai, Gang Xiong, Yong Zhao, and Longtao He. 2014. Analysis and Detection of Bogus Behavior in Web Crawler Measurement. *Procedia Computer Science* 31 (2014), 1084–1091.

[10] Kayleigh Barber. 2024. Future's Jon Steinberg shares his philosophy on AI content licensing deals — Digiday. https://digiday.com/podcasts/futures-jon-steinberg-shares-his-philosophy-on-ai-content-licensing-deals/.

[11] Muhammad Ahmad Bashir, Sajjad Arshad, Engin Kirda, William Robertson, and Christo Wilson. 2019. A Longitudinal Analysis of the ads.txt Standard. In *Proc. ACM Internet Measurement Conference 2019*. 294–307.

[12] Lucas Bellaiche, Rohin Shahi, Martin Harry Turpin, Anya Ragnhildstveit, Shawn Sprockett, Nathaniel Barr, Alexander Christensen, and Paul Seli. 2023. Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research: Principles and Implications* 8, 42 (2023), 22 pages.

[13] Alex Bocharov, Santiago Varagas, Adam Martinetti, Reid Tatoris, and Carlos Azevedo. 2024. Declare your AIndependence: block AI bots, scrapers and crawlers with a single click — Cloudflare. https://blog.cloudflare.com/declaring-your-aindependence-block-ai-bots-scrapers-and-crawlers-with-a-single-click/.

[14] Josiah D Boucher, Gillian Smith, and Yunus Doğan Telliel. 2024. Is Resistance Futile?: Early Career Game Developers, Generative AI, and Ethical Skepticism. In *Proc. of CHI Conference on Human Factors in Computing Systems 2024*. 1–13.

[15] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[16] Jack Brewster, Zach Fishman, and Isaiah Glick. 2024. AI Chatbots Are Blocked by 67% of Top News Sites, Relying Instead on Low-Quality Sources — News Guard. https://www.newsguardtech.com/special-reports/67-percent-of-top-news-sites-block-ai-chatbots/.

[17] Gordon Burtch, Dokyun Lee, and Zhichen Chen. 2023. The consequences of generative AI for UGC and online community engagement. *SSRN* (2023), 26 pages.

[18] Carbonmade. 2024. Terms and Conditions of Use. https://carbonmade.com/terms.

[19] Eva Cetinic and James She. 2022. Understanding and Creating Art with AI: Review and Outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 2 (2022), 1–22.

[20] Zi Chu, Steven Gianvecchio, and Haining Wang. 2018. Bot or Human? A Behavior-Based Online Bot Detection System. *From Database to Cyber Security: Essays Dedicated to Sushil Jajodia on the Occasion of His 70th Birthday* (2018), 432–449.

[21] Cloudflare. 2024. Verified Bots. https://radar.cloudflare.com/traffic/verified-bots.

[22] European Commission. 2024. First Draft of the General-Purpose AI Code of Practice published, written by independent experts. https://digital-strategy.ec.europa.eu/en/library/first-draft-general-purpose-ai-code-practice-published-written-independent-experts

[23] European Commission. 2024. The AI Act Explorer. https://artificialintelligenceact.eu/ai-act-explorer/.

[24] Common Crawl. 2025. Common Crawl — Open Repository of Web Crawl Data. https://commoncrawl.org/

[25] davepattern. 2024. DDoS from Anthropic AI. https://www.linode.com/community/questions/24842/ddos-from-anthropic-ai.

[26] deninho32. 2024. Claudebot attack. https://www.phpbb.com/community/viewtopic.php?t=2652265.

[27] Design and Artists Copyright Society (DACS). 2024. Artificial Intelligence and Artists' Work: A survey of artists on AI. https://cdn.dacs.org.uk/uploads/documents/News/Artificial-Intelligence-and-Artists-Work-DACS.pdf. (2024), 29 pages.

[28] Michael Dinzinger and Michael Granitzer. 2024. A Longitudinal Study of Content Control Mechanisms. In *Proc. of the ACM Web Conference*. 1382–1387.

[29] Michael Dinzinger, Florian Heß, and Michael Granitzer. 2024. A Survey of Web Content Control for Generative AI. (2024), 12 pages. arXiv:2404.02309 [cs.IR] https://arxiv.org/abs/2404.02309

[30] Ziv Epstein, Aaron Hertzmann, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R. Frank, Matthew Groh, Laura Herman, Neil Leach, Robert Mahari, Alex âĂIJSandyâĂİ Pentland, Olga Russakovsky, Hope Schroeder, and Amy Smith. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.

[31] Fairly Trained. 2024. Statement on AI training. https://www.aitrainingstatement.org/.

[32] Richard Fletcher. 2024. How many news websites block AI crawlers? *Reuters Institute Factsheets* (2024), 7 pages.

[33] Foundation Web Design & Development. 2022. What is DeviantArt's new "noai" and "noimageai" meta tag and how to install it. https://www.foundationwebdev.com/2022/11/noai-noimageai-meta-tag-how-to-install/.

[34] Sheera Frenkel and Stuart A. Thompson. 2023. Not for Machines to Harvest: Data Revolts Break Out Against A.I. — The New York Times. https://www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html

[35] Manuel B. Garcia. 2024. The Paradox of Artificial Creativity: Challenges and Opportunities of Generative AI Artistry. *Creativity Research Journal* (2024), 1–14.

[36] generosus. 2023. PSA | Bytedance and Bytespider Bots | Recommend Blocking. https://wordpress.org/support/topic/psa-bytedance-and-bytespider-bots-recommend-blocking/

[37] Jonathan Gillham. 2024. Block AI Bots from Crawling Websites Using Robots.txt — Originality.ai. https://originality.ai/ai-bot-blocking

[38] Google. 2024. Google Robots.txt Parser and Matcher Library. https://github.com/google/robotstxt.

[39] Paolo Grigis and Antonella De Angeli. 2024. Playwriting with Large Language Models: Perceived Features, Interaction Strategies and Outcomes. In *Proc. the International Conference on Advanced Visual Interfaces 2024*. 1–9.

[40] Alicia Guo, Shreya Sathyanarayanan, Leijie Wang, Jeffrey Heer, and Amy Zhang. 2025. From Pen to Prompt: How Creative Writers Integrate AI into their Writing Practice. (2025), 23 pages. arXiv:2411.03137 [cs.HC] https://arxiv.org/abs/2411.03137

[41] Eszter Hargittai. 2009. An Update on Survey Measures of Web-Oriented Digital Literacy. *Social Science Computer Review* 27, 1 (2009), 130–137.

[42] Joo-Wha Hong and Nathaniel Ming Curran. 2019. Artificial Intelligence, Artists, and Art: Attitudes Toward Artwork Produced by Humans vs. Artificial Intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 2s (2019), 1–16.

[43] Xinyi Hou, Yanjie Zhao, and Haoyu Wang. 2024. On the (In)Security of LLM App Stores. (2024), 17 pages. arXiv:2407.08422 [cs.CR] https://arxiv.org/abs/2407.08422

[44] Hongxian Huang, Runshan Fu, and Anindya Ghose. 2023. Generative AI and Content Creators: Evidence from Digital Art Platforms. *SSRN* (2023), 41 pages.

[45] Christos Iliou, Theodoros Kostoulas, Theodora Tsikrika, Vasilis Katos, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2021. Detection of Advanced Web Bots by Combining Web Logs with Mouse Behavioural Biometrics. *Digital Threats: Research and Practice* 2, 3 (2021), 1–26.

[46] Christos Iliou, Theodoros Kostoulas, Theodora Tsikrika, Vasilis Katos, Stefanos Vrochidis, and Yiannis Kompatsiaris. 2019. Towards a framework for detecting advanced web bots. In *Proc. of 14th International Conference on Availability, Reliability and Security*. 1–10.

[47] Katy Ilonka Gero, Meera Desai, Carly Schnitzler, Nayun Eom, Jack Cushman, and Elena L. Glassman. 2024. Creative Writers' Attitudes on Writing as Training Data for Large Language Models. In *Proc. of CHI Conference on Human Factors in Computing Systems 2025*. 1–16.

[48] Imperva. 2024. 2024 Bad Bot Report. https://www.imperva.com/resources/resource-library/reports/2024-bad-bot-report/.

[49] Gregoire Jacob, Engin Kirda, Christopher Kruegel, and Giovanni Vigna. 2012. PUBCRAWL: Protecting Users and Businesses from CRAWLers. In *Proc. of 21st USENIX Security Symposium*. 507–522.

[50] Steve TK Jan, Qingying Hao, Tianrui Hu, Jiameng Pu, Sonal Oswal, Gang Wang, and Bimal Viswanath. 2020. Throwing Darts in the Dark? Detecting Bots with Limited Data using Neural Data Augmentation. In *Proc. of 2020 IEEE Symposium on Security and Privacy*. IEEE, 1190–1206.

[51] Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. AI Art and its Impact on Artists. In *Proc. of AAAI ACM Conference on AI, Ethics, and Society 2023*. 363–374.

[52] Hannah Johnston and David Thue. 2024. Understanding Visual Artists' Values and Attitudes towards Collaboration, Technology, and AI. In *Proc. 50th Graphics Interface Conference*. 1–9.

[53] Ben Jones, Tzu-Wen Lee, Nick Feamster, and Phillipa Gill. 2014. Automated Detection and Fingerprinting of Censorship Block Pages. In *Proc. of ACM Internet Measurement Conference 2014*. 299–304.

[54] Hugo Jonker, Benjamin Krumnow, and Gabry Vlot. 2019. Fingerprint Surface-Based Detection of Web Bot Detectors. In *Proc. of 24th European Symposium on Research in Computer Security*. Springer, 586–605.

[55] Reishiro Kawakami and Sukrit Venkatagiri. 2024. The Impact of Generative AI on Artists. In *Proc. of 16th Conference on Creativity & Cognition*. 79–82.

[56] Paul Keller. 2023. Protecting Creatives or Impeding Progress? Machine learning and the EU copyright framework — Open Future. https://openfuture.eu/blog/protecting-creatives-or-impeding-progress/.

[57] Kate Knibbs. 2024. CondÃĬ Nast Signs Deal With OpenAI — Wired. https://www.wired.com/story/conde-nast-openai-deal/.

[58] Kate Knibbs. 2024. The Race to Block OpenAI's Scraping Bots Is Slowing Down — Wired. https://www.wired.com/story/open-ai-publisher-deals-scraping-bots/.

[59] Robb Knight. 2024. Perplexity AI Is Lying about Their User Agent. https://rknight.me/blog/perplexity-ai-is-lying-about-its-user-agent/

[60] Santanu Kolay, Paolo D'Alberto, Ali Dasdan, and Arnab Bhattacharjee. 2008. A Larger Scale Study of Robots.txt. In *Proc. of 17th World Wide Web Conference*. 1171–1172.

[61] Martijn Koster, Gary Illyes, Henner Zeller, and Lizzi Sassman. 2022. Robots Exclusion Protocol. RFC 9309. https://doi.org/10.17487/RFC9309

[62] Shinil Kwon, Young-Gab Kim, and Sungdeok Cha. 2012. Web robot detection based on pattern-matching technique. *Journal of Information Science* 38, 2 (2012), 118–126.

[63] IAB Tech Lab. 2025. Global Privacy Protocol. https://iabtechlab.com/gpp/

[64] Rita Latikka, Jenna Bergdahl, Nina Savela, and Atte Oksanen. 2023. AI as an Artist? A Two-Wave Survey Study on Attitudes Toward Using Artificial Intelligence in Art. *Poetics* 101 (2023), 11 pages.

[65] Junsup Lee, Sungdeok Cha, Dongkun Lee, and Hyungkyu Lee. 2009. Classification of web robots: an empirical study based on over one billion requests. *Computers & Security* 28, 8 (2009), 795–802.

[66] Jie Li, Hancheng Cao, Laura Lin, Youyang Hou, Ruihao Zhu, and Abdallah El Ali. 2024. User Experience Design Professionals' Perceptions of Generative Artificial Intelligence. In *Proc. of CHI Conference on Human Factors in Computing Systems 2024*. 1–18.

[67] Xigao Li, Babak Amin Azad, Amir Rahmati, and Nick Nikiforakis. 2021. Good Bot, Bad Bot: Characterizing Automated Browsing Activity. In *Proc. of 2021 IEEE Symposium on Security and Privacy*. IEEE, 1589–1605.

[68] Baker & Hostetler LLP. 2024. Case Tracker: Artificial Intelligence, Copyrights and Class Actions. https://www.bakerlaw.com/services/artificial-intelligence-ai/case-tracker-artificial-intelligence-copyrights-and-class-actions/

[69] Joseph Saveri Law Firm LLP. 2023. Class Action Filed Against Stability AI, Midjourney, and DeviantArt for DMCA Violations, Right of Publicity Violations, Unlawful Competition, Breach of TOS. https://www.prnewswire.com/news-releases/class-action-filed-against-stability-ai-midjourney-and-deviantart-for-dmca-violations-right-of-publicity-violations-unlawful-competition-breach-of-tos-301721869.html

[70] Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Anis, An Dinh, Caroline Chitongo, Da Yin, Damien Sileo, Deividas Mataciunas, Diganta Misra, Emad Alghamdi, Enrico Shippole, Jianguo Zhang, Joanna Materzynska, Kun Qian, Kush Tiwary, Lester Miranda, Manan Dey, Minnie Liang, Mohammed Hamdy, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Shrestha Mohanty, Vipul Gupta, Vivek Sharma, Vu Minh Chien, Xuhui Zhou, Yizhi Li, Caiming Xiong, Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, Jad Kabbara, and Sandy Pentland. 2024. Consent in Crisis: The Rapid Decline of the AI Data Commons. (2024), 41 pages. arXiv:2407.14933 [cs.CL] https://arxiv.org/abs/2407.14933

[71] Anália G. Lourenço and Orlando O. Belo. 2006. Catching Web Crawlers in the Act. In *Proc. of 6th International Conference on Web Engineering*. 265–272.

[72] Juniper Lovato, Julia Witte Zimmerman, Isabelle Smith, Peter Dodds, and Jennifer L. Karson. 2024. Foregrounding Artist Opinions: A Survey Study on Transparency, Ownership, and Fairness in AI Generative Art. In *Proc. of AAAI ACM Conference on AI, Ethics, and Society 2024*. 905–916.

[73] Abhijeeth Madhu. 2023. Survey Reveals 9 out of 10 Artists Believe Current Copyright Laws are Outdated in the Age of Generative AI Technology — Book An Artist Blog. https://bookanartist.co/blog/2023-artists-survey-on-ai-technology/

[74] Bron Maher. 2024. Revealed: Which of the top 100 UK and US news websites are blocking AI crawlers — PressGazette. https://pressgazette.co.uk/platforms/news-sites-block-ai-web-crawlers-chatgpt-google/

[75] Meta. 2024. Meta Web Crawlers. https://developers.facebook.com/docs/sharing/webmasters/web-crawlers/.

[76] Elzė Sigutė Mikalonytė and Markus Kneer. 2022. Can Artificial Intelligence Make Art?: Folk Intuitions as to whether AI-driven Robots Can Be Viewed as Artists and Produce Art. *ACM Transactions on Human-Robot Interaction* 11, 4 (2022), 1–19.

[77] Cullen Miller. 2023. ai.txt: A new way for websites to set permissions for AI — Spawning. https://spawning.substack.com/p/aitxt-a-new-way-for-websites-to-set.

[78] Piotr Mirowski, Juliette Love, Kory Mathewson, and Shakir Mohamed. 2024. A Robot Walks into a Bar: Can Language Models Serve as Creativity Support Tools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians. In *Proc. of the ACM Conference on Fairness, Accountability, and Transparency 2024*. 1622–1636.

[79] Martin Monperrus. 2024. crawler-user-agents. https://github.com/monperrus/crawler-user-agents.

[80] Alexis Newton and Kaustubh Dhole. 2023. Is AI Art Another Industrial Revolution in the Making? (2023), 6 pages. arXiv:2301.05133 [cs.AI] https://arxiv.org/abs/2301.05133

[81] Republic of Singapore. 2021. Singapore Copyright Act of 2021: Section 244 (English Translation). https://sso.agc.gov.sg/Acts-Supp/22-2021/Published/?ProvIds=pr244-.

[82] OpenAI. 2023. Our approach to AI safety. https://openai.com/index/our-approach-to-ai-safety/

[83] Karla Ortiz. 2024. Why AI Models are not inspired like humans. https://www.kortizblog.com/blog/why-ai-models-are-not-inspired-like-humans

[84] Stack Overflow. 2024. Stack Overflow and OpenAI Partner to Strengthen the World's Most Popular Large Language Models. https://stackoverflow.co/company/press/archive/openai-partnership.

[85] palewire. 2025. Who blocks OpenAI, Google AI and Common Crawl? https://palewi.re/docs/news-homepages/openai-gptbot-robotstxt.html

[86] Sungjin Park. 2024. The work of art in the age of generative AI: aura, liberation, and democratization. *AI & Society* (2024), 1–10.

[87] Perplexity. 2025. Perplexity Crawlers. https://docs.perplexity.ai/guides/bots.

[88] Kien Pham, Aécio Santos, and Juliana Freire. 2016. Understanding Website Behavior based on User Agent. In *Proc. of 39th ACM SIGIR Conference on Research and Development in Information Retrieval*. 1053–1056.

[89] Tara Poteat and Frank Li. 2021. Who You Gonna Call? An Empirical Evaluation of Website security.txt Deployment. In *Proc. of ACM Internet Measurement Conference 2021*. 526–532.

[90] Heila Precel, Allison McDonald, Brent Hecht, and Nicholas Vincent. 2024. A Canary in the AI Coal Mine: American Jews May Be Disproportionately Harmed by Intellectual Property Dispossession in Large Language Model Training. In *Proc. of CHI Conference on Human Factors in Computing Systems 2024*. 1–17.

[91] PRNewswire. 2024. Dotdash Meredith Announces Strategic Partnership with OpenAI, Bringing Iconic Brands and Trusted Content to ChatGPT. https://dotdashmeredith.mediaroom.com/2024-05-07-Dotdash-Meredith-Announces-Strategic-Partnership-with-OpenAI,-Bringing-Iconic-Brands-and-Trusted-Content-to-ChatGPT.

[92] Gayatri Raman and Erin Brady. 2024. Exploring Use and Perceptions of Generative AI Art Tools by Blind Artists. (2024), 4 pages. arXiv:2409.08226 [cs.HC] https://arxiv.org/abs/2409.08226

[93] rejeptai. 2024. Why doesn't ClaudeBot/Anthropic obey robots.txt? https://www.reddit.com/r/Anthropic/comments/1c8tu5u/why_doesnt_claudebot_anthropic_obey_robotstxt/

[94] Copyright Research and Information Center. 2019. Copyright Law of Japan: Chapter II Rights of Authors (English Translation). https://www.cric.or.jp/english/clj/cl2.html.

[95] Stefano Rovetta, Alberto Cabri, Francesco Masulli, and Grażyna Suchacka. 2019. Bot or Not? A Case Study on Bot Recognition from Web Session Logs. *Quantifying and Processing Biomedical and Behavioral Signals* (2019), 197–206.

[96] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. In *Proc. of ACM Internet Measurement Conference 2018*. 478–493.

[97] M. H. M. Schellekens. 2013. Robot.txt: balancing interests of content producers and content users. *Bridging Distances in Technology and Regulation* (2013), 173–187.

[98] Barry Schwartz. 2023. Google-Extended does not stop Google Search Generative Experience from using your site's content. https://searchengineland.com/google-extended-does-not-stop-google-search-generative-experience-from-using-your-sites-content-433058.

[99] Yoast SEO. 2025. SEO starts with Yoast. https://yoast.com/

[100] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. 2023. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models. In *Proc. of 32nd USENIX Security Symposium*. 2187–2204.

[101] Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. 2024. Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. In *Proc. of IEEE Symposium on Security and Privacy 2024*. 807–825.

[102] Jingyu Shi, Rahul Jain, Runlin Duan, and Karthik Ramani. 2023. Understanding Generative AI in Art: An Interview Study with Artists on G-AI from an HCI

Perspective. (2023), 15 pages. arXiv:2310.13149 [cs.HC] https://arxiv.org/abs/2310.13149
[103] Dusan Stevanovic, Aijun An, and Natalija Vlajic. 2012. Feature evaluation for web crawler detection with data mining techniques. *Expert Systems with Applications* 39, 10 (2012), 8707–8717.
[104] Dongxun Su, Yanjie Zhao, Xinyi Hou, Shenao Wang, and Haoyu Wang. 2024. GPT Store Mining and Analysis. (2024), 16 pages. arXiv:2405.10210 [cs.LG] https://arxiv.org/abs/2405.10210
[105] Grażyna Suchacka, Alberto Cabri, Stefano Rovetta, and Francesco Masulli. 2021. Efficient on-the-fly Web bot detection. *Knowledge-Based Systems* 223 (2021), 16 pages.
[106] Mark Sullivan. 2024. AI Companies Ignoring Robots.txt. https://mjtsai.com/blog/2024/06/24/ai-companies-ignoring-robots-txt/
[107] Yang Sun, Ziming Zhuang, Isaac G. Councill, and C. Lee Giles. 2007. Determining Bias to Search Engines from Robots.txt. In *Proc. of IEEE WIC ACM Web Intelligence and Intelligent Agent Technology 2007*. 149–155.
[108] Yang Sun, Ziming Zhuang, and C. Lee Giles. 2007. A Large-Scale Study of Robots.txt. In *Proc. of 16th World Wide Web Conference*. 1123–1124.
[109] David SÃĺnÃĺcal. 2024. The Web Scraping Problem: Part 1 — Akamai. https://www.akamai.com/blog/security/the-web-scraping-problem-part-1.
[110] Reid Tatoris, Harsh Saxena, and Luis Miglietti. 2025. Trapping misbehaving bots in an AI Labyrinth — Cloudflare. https://blog.cloudflare.com/ai-labyrinth/
[111] Antoine Vastel, Walter Rudametkin, Romain Rouvoy, and Xavier Blanc. 2020. FP-Crawlers: Studying the Resilience of Browser Fingerprinting to Block Crawlers. In *Proc. of NDSS Workshop on Measurements, Attacks, and Defenses for the Web*. 13 pages.
[112] James Vincent. 2023. Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content — The Verge. https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit
[113] Dark Visitors. 2024. Agents. https://darkvisitors.com/agents.
[114] Dark Visitors. 2025. Track the AI Agents and Bots Crawling Your Website. https://darkvisitors.com/
[115] W3Techs. 2024. Usage statistics and market shares of reverse proxy services. https://w3techs.com/technologies/overview/proxy.
[116] Wikipedia contributors. 2020. Global Privacy Control — Wikipedia. https://en.wikipedia.org/wiki/Global_Privacy_Control
[117] Wix. 2025. Wix.com Terms of Use. https://www.wix.com/about/terms-of-use.
[118] Chloe Xiang. 2022. Artists Are Revolting Against AI Art on ArtStation — Vice. https://www.vice.com/en/article/ake9me/artists-are-revolt-against-ai-art-on-artstation
[119] Chyan Yang and Hsien-Jyh Liao. 2010. Using the Robots. txt and Robots Meta tags to implement online copyright and a related amendment. *Library Hi Tech* 28, 1 (2010), 94–106.
[120] Zejun Zhang, Li Zhang, Xin Yuan, Anlan Zhang, Mengwei Xu, and Feng Qian. 2024. A First Look at GPT Apps: Landscape and Vulnerability. (2024), 11 pages. arXiv:2402.15105 [cs.CR] https://arxiv.org/abs/2402.15105
[121] Eric Zhou and Dokyun Lee. 2024. Generative artificial intelligence, human creativity, and art. *PNAS Nexus* 3, 3 (2024), 8 pages.
[122] Viola Zhou. 2023. AI is already taking video game illustrators' jobs in China — Rest of World. https://restofworld.org/2023/ai-china-video-game-layoffs-illustrators/

A Ethics

We believe our work has very low ethical risk. Our user study is approved by the IRB at our institution. Our longitudinal analysis leverages common crawl data, which is publicly available and does not contain any personal information, and our active blocking experiments are conducted at a responsible rate. We also make our data and code available to the community at https://github.com/ucsdsysnet/ai-crawler-imc-25.

B Historic Use of Robots.txt

B.1 Common Crawl Snapshots

For our historic robots.txt analysis (Section 3), we used data from 15 consecutive snapshots from Common Crawl from October 2022 to October 2024. Table 3 lists each Common Crawl snapshot, the months it covers (as reported by Common Crawl's website), and the number of sites that are in the Stable Top 100K and have a robots.txt file in each particular snapshot. For each snapshot, Common Crawl may crawl a site several times over the period in which the data for the snapshot was collected. In these cases, we deduplicate the robots.txt files by taking the most recent non-errored crawl in the snapshot. The Common Crawl crawler also does not follow redirects. To improve our coverage, for domains that returned a non-200 HTTP status code to Common Crawl (such as 301 Redirect), we also checked Common Crawl for the robots.txt file for the domain prepended with "www." (if not already) and without (if already prepended).

Snapshot	Month	# Sites	+ robots.txt
2022-05	Sep/Oct 2022	40177	31494
2022-21	Nov/Dec 2022	40614	31536
2022-40	Jan/Feb 2023	39080	30063
2023-06	Mar/Apr 2023	39216	29963
2023-14	May/Jun 2023	39212	30107
2023-23	Sep/Oct 2023	39033	29721
2023-40	Nov/Dec 2023	39722	30060
2023-50	Feb/Mar 2024	41446	31282
2024-10	Apr 2024	41640	31010
2024-18	May 2024	41004	30763
2024-22	Jun 2024	41047	30661
2024-26	Jul 2024	40927	30526
2024-33	Aug 2024	40455	29922
2024-38	Sep 2024	40444	29806
2024-42	Oct 2024	40420	29867

Table 3: Snapshots used in the historic AI crawler analysis: the months they cover, the number of sites in the Stable Top 100K in each snapshot, and the number of those sites that have a robots.txt file in the snapshot.

B.2 Robots.txt edge cases

When experimenting with robots.txt parsers from both Google and [70], we discovered three edge cases that can lead to very different interpretations of a robots.txt file depending on whether a parser is fully compliant with RFC 9309.

Case 1. For the following robots.txt, a compliant parser will ignore comments or newlines after the "User-agent" line and respect the "Disallow" directives. If a parser does not handle such comments or newlines correctly, the parser may skip and ignore the "Disallow" directives:

```
User-agent: *
# Blog restrictions
Disallow: /blog/latest/*
Disallow: /blogs/*
```

Case 2. RFC 9309 allows "User-agent" directives to be grouped as shown below. A non-compliant parser, however, can ignore all such grouped "User-agent" lines except for the last when parsing robots.txt:

```
User-agent: GPTBot
User-agent: anthropic-ai
User-agent: Claudebot
```

```
Disallow: /
```

Case 3. Using unsupported directives can have unintended consequences. For example, "Crawl-delay" is a non-standard extension supported by some crawlers and ignored by others, a situation that can lead to unexpected results depending on the parser used by the crawler. Google's compliant robots.txt parser will ignore the "Crawl-delay" directive and effectively treat it as a blank line. As a result, in the following robots.txt the "User-agent: *" directive will be combined with the "User-agent: GoogleBot" directive due to the grouping rule (ignoring "Crawl-delay" and effectively grouping the two "User-agent" lines together):

```
User-agent: *
Disallow: /

User-agent: *
Crawl-delay: 5

User-agent: GoogleBot
Allow: /
Disallow: /z/
```

In contrast, a parser that obeys the non-standard "Crawl-delay" directive will not group together the two "User-agent" lines (only the GoogleBot user agent will be associated with the two "Allow/Disallow" rules).

B.3 Domains that explicitly allow GPTBot

Table 4 shows the list of domains that explicitly and fully allow GPTBot in their robots.txt with a directive such as:

```
User-agent: GPTBot
Allow: /
```

as well as the Common Crawl snapshot in which we first observed this behavior. We note that five sites (nfhs.org, 10best.com, ground.news, network54.com, and tarleton.edu) have persistently allowed GPTBot since around the time of its release to our latest snapshot.

C Active Blocking

C.1 Squarespace Restricted AI Bots

The following directives are added to the robots.txt file for a Squarespace site when a customer turns off the "Artificial Intelligence Crawlers" option:

```
User-agent: GPTBot
User-agent: ChatGPT-User
User-agent: CCBot
User-agent: anthropic-ai
User-agent: Google-Extended
User-agent: FacebookBot
User-agent: Claude-Web
User-agent: cohere-ai
User-agent: PerplexityBot
User-agent: Applebot-Extended
```

Site	Snapshot	Site	Snapshot
nfhs.org	2023-40	bleedcubbieblue.com	2024-42
10best.com	2023-40	popsugar.com	2024-42
ground.news	2023-40	voxmedia.com	2024-42
opindia.com	2024-42	patspulpit.com	2024-42
tarleton.edu	2023-50	barcablaugranes.com	2024-42
alldatasheet.com	2024-42	eater.com	2024-42
bestproductsreviews.com	2024-42	popsugar.co.uk	2024-42
network54.com	2023-50	prideofdetroit.com	2024-42
care.com	2024-42	royalsreview.com	2024-42
kbs.co.kr	2024-42	truebluela.com	2024-42
brit.co	2024-42	thrillist.com	2024-42
lonza.com	2024-42	sbnation.com	2024-42
millersville.edu	2024-42	arrowheadpride.com	2024-42
icelandair.com	2024-42	theringer.com	2024-42
customink.com	2024-42	adslzone.net	2024-42
celebmafia.com	2024-18	milehighreport.com	2024-42
credit-agricole.fr	2024-42	polygon.com	2024-42
adelaidenow.com.au	2024-42	racked.com	2024-42
dailytelegraph.com.au	2024-42	behindthesteelcurtain.com	2024-42
walkhighlands.co.uk	2024-42	bavarianfootballworks.com	2024-42
softonic-ar.com	2024-22	bleedinggreennation.com	2024-42
heraldsun.com.au	2024-42	silverscreenandroll.com	2024-42
royalsocietypublishing.org	2024-22	gnc.com	2024-42
softonic.com	2024-42	cagesideseats.com	2024-42
shopstyle.com	2024-42	blazersedge.com	2024-42
couriermail.com.au	2024-42	badlefthook.com	2024-42
theaustralian.com.au	2024-42	cincyjungle.com	2024-42
news.com.au	2024-42	hogshaven.com	2024-42
kaufland.de	2024-42	bigblueview.com	2024-42
sendpulse.com	2024-26	ninersnation.com	2024-42
washingtonexaminer.com	2024-33	pinstripealley.com	2024-42
thedodo.com	2024-42	bloggingtheboys.com	2024-42
g2a.com	2024-42	quickbase.com	2024-42
fieldgulls.com	2024-42	embluemail.com	2024-42
recode.net	2024-42	softonic.com.br	2024-42
novartis.com	2024-38	stimulustech.com	2024-42
mmafighting.com	2024-42	searchenginejournal.com	2024-42
vox.com	2024-42	giant-bicycles.com	2024-42
mmamania.com	2024-42	realself.com	2024-42

Table 4: Domains that explicitly and fully allow GPTBot in their robots.txt, and the Common Crawl snapshot in which we first observed this behavior.

```
Disallow: /
```

C.2 Cloudflare "Definitely Automated"

The following list shows the user agents we inferred Cloudflare's "Definitely Automated" setting to block:

```
360Spider
AHC
aiohttp
anthropic-ai
Apache-HttpClient
axios
binlar
Bytespider
CCBot
centurybot
Claudebot
curl
Diffbot
Go-http-client
grub.org
HeadlessChrome
httpx
libwww-perl
magpie-crawler
MeltwaterNews
node-fetch
Nutch
omgili
PerplexityBot
PhantomJS
PHP-Curl-Class
PiplBot
python-requests
Python-urllib
Scrapy
serpstatbot
Teoma
W3C-checklink
wget
```

We note that IP address likely plays a role in the operation of this setting to block "fake" verified bots (e.g., a request that claims to be a particular Cloudflare Verified Bot, but does not come from a documented IP address). We exclude these user agents from the list, but note that the list of Cloudflare verified bots is publicly available [21].

C.3 Cloudflare's "Block AI Scrapers and Crawlers"

The following user agents are blocked by Cloudflare's "Block AI Scrapers and Crawlers" option:

`Amazonbot`	`Diffbot/`
`AwarioRssBot`	`GPTBot/`
`AwarioSmartBot`	`magpie-crawler`
`Bytespider`	`MeltwaterNews`
`CCBot/`	`omgili/`
`ChatGPT-User`	`PerplexityBot`
`Claude-Web`	`PiplBot`
`ClaudeBot`	`YouBot`
`cohere-ai`	

Note that AwarioRssBot, AwarioSmartBot, magpie-crawler, and MeltwaterNews are not in the Dark Visitors list of AI user agents.

D Artist Survey

D.1 Survey Questions

In this section, we provide the list of questions that we asked in the artist survey. We omit the questions related to contact information and compensation. Our study was approved by our university's Institutional Review Board (IRB).

Questions about artistic background

Q1. Do you consider yourself a professional artist?
- Yes • No

Q2. What portion of your income comes from your art?
- I haven't made any money from my art
- I make some income from my art but it's not the main source
- My art is my main source of income

Q3. How long have you been making money from your art?
- Less than 1 year • 1–5 years • 5–10 years • 10 years or more

Q4. What type of art do you do? (Select all that apply)
- Concept Art • Traditional Painting and Drawing • Photography
- Abstract • Illustration • Game Art • Anime and Manga Art
- Digital 2D • Digital 3D • Traditional Sculpting • Environmental
- Character and Creature Design • Comicbook Art • Matte Painting
- Items Props • Other (please specify)

Q5. Which country do you live in?
- Australia • Brazil • Canada • China • France • Germany • India
- Italy • Japan • Mexico • Russia • South Africa • Spain
- United Kingdom • United States • Other (please specify)

Questions about technical background

Q6. How familiar are you with the following computer and internet items? *(1-5; 1 = no understanding, 5 = full understanding.)*
- Website • Generative AI • Search engine
- Nearest diffusion tree • Robots.txt

Q7. Do you post your art online?
- Yes • No

Q8. Where do you post art online? (Select all that apply)
- Social Media (Instagram, LinkedIn, …)
- Art Platforms (ArtStation, DeviantArt, …)
- Personal Website
- Art Seller Websites (Artsy, Artrepreneur, …)
- Other (please specify)

Q9. How do you host your personal website?
- I have my own server • Free service (e.g., free server with AWS)
- Paid service (e.g., Squarespace with a custom domain)
- Other (please specify)

Q10. What is the name of the service you use?
Answer: _____

Q11. Why did you choose the service?
Answer: _____

Q12. [Optional] If you're comfortable, please share a link to your personal website.
Answer: _____

Questions about impressions of AI art and their actions

Q13. How familiar are you with AI-generated art?
- Not familiar at all
- Slightly familiar
- Somewhat familiar
- Moderately familiar
- Very familiar

Q14. Do you use AI in your artistic process?
- Never • Rarely • Sometimes • Often • Always

Q15. Please briefly describe your general impression of AI-generated art.
Answer: _____

Q16. How much impact do you expect AI-generated art to have on your job security?
- No impact • Minor impact • Moderate impact
- Significant impact • Severe impact

Q17. Have you taken any actions because of the increasing use of AI-generated art in recent years?
- Yes • No

Q18. What actions have you taken? (Select all that apply)
- Reducing the amount of my artwork that I share online
- Actively removing my old artwork from the Internet
- Posting lower resolution versions of my artwork online
- Learning about AI art tools and possibly using them
- Preventing my websites from being scraped
- Using Glaze to protect my art before posting
- Other (please specify)

Q19. Please elaborate on how you prevent your websites from being scraped.
Answer: _____

Q20. Do you plan to take any actions because of the increasing use of AI-generated art in recent years?
- Yes • No

Q21. What actions do you plan to take? (Select all that apply)
- Reducing the amount of my artwork that I share online
- Actively removing my old artwork from the Internet
- Posting lower resolution versions of my artwork online
- Learning about AI art tools and possibly using them
- Using Glaze to protect my art before posting
- Preventing my websites from being scraped
- Other (please specify)

Q22. If your website hosting platform offers a mechanism (e.g. by clicking a button) to tell AI companies that you would like them not to scrape your website, how likely will you enable this mechanism?

• Not likely at all • Unlikely • Neutral / Undecided
• Likely • Very likely
Why or why not? Answer: _____

Q23. If your website hosting platform offers a mechanism (e.g. by clicking a button) to block AI companies from scraping your website, how likely will you enable this mechanism?
• Not likely at all • Unlikely • Neutral / Undecided
• Likely • Very likely
Why or why not? Answer: _____

Questions about knowledge of robots.txt

Q24. Have you heard about robots.txt before today?
• Yes • No

Description of robots.txt for artists who select "no" in Q24. This description is generated with the help of ChatGPT.

Do you know that over 90% of artists don't realize they can use a simple tool called robots.txt to stop automated programs (also known as bots) from downloading content from their websites? Think of robots.txt as a "Do Not Enter" sign for automated programs that browse the internet. When placed on a website, it tells these automated programs which parts of the site they're not allowed to access. While it won't stop every bot, it works like a polite request to keep things like personal galleries or portfolios hidden from search engines or unwanted bots. This is an easy way for artists to protect their work and control how it appears online, without needing to dive into complicated tech or legal steps. Adding a robots.txt file can be a quick win for maintaining privacy and keeping unwanted eyes off your art.

That being said, it is important to note that not all companies respect robots.txt—some may ignore it entirely if they choose to.

Q25. Briefly describe what you think robots.txt does.
Answer: _____

Q26. Would you consider adopting robots.txt in the future?
• Not likely at all • Unlikely • Neutral / Undecided
• Likely • Very likely
Why or why not? *(Open-ended)*

Q27. Robots.txt is a standardized way to declare "do not crawl," and most companies respect it. How likely do you think AI companies will respect robots.txt?
• Not likely at all • Unlikely • Neutral / Undecided
• Likely • Very likely
Why or why not? Answer: _____

Q28. Have you checked the robots.txt of websites where you post your work?
• Yes • No

Q29. Can you control (edit or modify) the content of the robots.txt of websites where you post your work?
• I have full control over the full content of robots.txt
• I can click some buttons to switch between a few presets
• I have no control over the content
• I am not sure
• Other (please specify)

Q30. How did you get the current content of robots.txt?
• Provided by my website hosting platform
• Copied from the Internet (e.g., a blog)
• Created my own robots.txt
• Other (please specify)

Q31. Do you currently use robots.txt to disallow bots from AI companies from scraping websites where you post your art?
• Yes • No
Why? Answer: _____
Why not?
• I am concerned it will impact the discoverability of my website online
• I don't mind AI training on my art
• I don't know how to do it
• Other (please specify)

Q32. [Optional] Do you face any obstacles in adopting robots.txt? (Select all that apply)
• I have trouble finding how to edit the robots.txt
• I find it hard to write the robots.txt
• I don't know how to use it
• Other (please specify)

D.2 Demographics

This section presents the demographics of the participants in our survey. As previously mentioned, we focus on their artistic background, as it is the most relevant to our study.

Duration	Count
Less than 1 year	17
1–5 years	68
5–10 years	44
10 years or more	47
Total	**176**

Table 5: How long participants have been making money from their art.

Table 5 presents a breakdown of how long participants have been making money from their art. The majority of respondents (68) have been doing so for 1–5 years, whereas only 17 have been making money from their art for less than a year. Over half of the respondents (91) have been making money from their art for at least 5 years.

Continent	Count
North America	109
Europe	52
Asia	21
South America	18
Africa	2
Oceania	1
Total	**203**

Table 6: Continent of residence of participants.

Table 6 presents a breakdown of the continent of residence of participants. The majority of participants (109) are from North America, with 89 of them from the United States. The second largest group is from Europe (52), with 18 from the United Kingdom, five from Poland, and another five from Germany. The third largest group is from Asia (21), with nine from The Philippines. The remaining participants are from South America (18), Africa (2), and Oceania (1).

Table 7 presents a breakdown of the top five types of art participants do. Each participant can select every type of art they do, so the total number of responses is greater than the number of participants. The most common type of art is illustration (163), followed

Art Type	Count
Illustration	163
Digital 2D	143
Character and Creature Design	99
Traditional Painting and Drawing	78
Concept Art	68
Total	**551**

Table 7: Top five types of art participants do.

by digital 2D (143), character and creature design (99), traditional painting and drawing (78), and concept art (68).

Term	Average Familiarity
Website	4.60
Search Engine	4.35
Generative AI	3.89
Robots.txt	1.99
Nearest diffusion tree	1.56

Table 8: Participant's average familiarity with various terms. The average is on a scale from 1 to 5, where 1 represents no understanding and 5 represents full understanding. Following the work of Hargittai [41], we also include a bogus item "Nearest diffusion tree", indicated in italics.

Table 8 presents our participant's average familiarity with various terms. This question is designed to assess our participants's digital literacy. The average is on a scale from 1 to 5, where 1 represents no understanding and 5 represents full understanding. Following the work of Hargittai [41], we also include a bogus item "Nearest diffusion tree", indicated in italics. The most familiar term is "website" (4.60), followed by "search engine" (4.35), "generative AI" (3.89), and "robots.txt" (1.99). The least familiar term is "nearest diffusion tree" (1.56). Given that this bogus term was rated as the lowest compared to the other four terms, we conclude that our participants do not select randomly. This data also suggests that our participants are relatively familiar with general terms such as "website", "search engine", and "generative AI", but much less familiar with "robots.txt". This result is consistent with our other findings in Section 4.

D.3 Codebook

This section details the codebook we used to analyze the qualitative data collected from artists. Specifically, Table 9 lists other actions taken by artists in response to AI-generated art; Table 10 lists reasons why artists would not adopt robots.txt; Table 11 lists reasons why artists would enable a mechanism that blocks AI crawlers; and Table 12 lists reasons why artists do not trust AI companies to respect robots.txt.

Theme	Description	Example
Modify post	Artists alter the content or format of the artwork they share online.	"Overlaying watermarks or art filters to modify the artwork"
Switch platforms	Artists migrate to alternative sites or remove their work from certain platforms.	"Use Cara instead of Instagram"
Raise awareness	Artists publicly highlight issues affecting them or the community.	"Spreading awareness about the damage AI-generated art does"
Unionize	Artists organize collectively to negotiate or advocate for shared interests.	"Connecting with groups of professional artists being impacted to search for collective solutions for our field"
Change career path	Artists pivot to a different professional direction.	"I left school and am taking a gap year to reevaluate my life"
Miscellaneous	Additional strategies not covered above.	"Using block lists to block AI art accounts"

Table 9: Codebook for other actions taken by artists in response to AI-generated art.

Theme	Description	Example
Efficacy	Artists are concerned about the efficacy of robots.txt given its voluntary nature.	"if the companies can ignore it why would they respect it considering what they already do"
Usability	Artists are concerned about the complexity of implementing or using robots.txt.	"It sounds like something difficult to use"
More information	Artists want to gather more information about robots.txt before making a decision.	"Not informed enough about it"
No personal website	Artists do not have a personal website.	"I do not have a personal website"
Search results	Artists are concerned about robots.txt impacting the search results of their websites.	"If it hides things from *search engines* then how will people find my work?..."

Table 10: Codebook for why artists would not adopt robots.txt.

Theme	Description	Example
Protection	Artists want to protect their work.	"To protect my original concepts and visual brand (aka original character designs and artstyle)"
Consent	Artists do not want their work to be crawled and do not consent to crawling.	"I havent given AI companies permission to use my work"
Compensation	Artists are not compensated while AI companies profit from their work.	"..., and I do not want other companies to profit off of it without my knowledge, permission, or without fair compensation towards the source."
Useful mechanism	Artists see this mechanism as useful and reassuring.	"Adds a sense of security and ease of use."
Legal benefit	Artists believe such mechanisms could be potentially useful in legal cases.	"..., it is a measure to reinforce a statement that we do not condone with these practices and will probably benefit in a possible lawsuit in the future."
Misc	Additional reasons not covered above.	"At this point if the option is presented I'll do my research on it and if it seems legitimate I'll do it on principle."

Table 11: Codebook for why artists would enable a mechanism that blocks AI crawlers.

Theme	Description	Example
Track record	AI companies have a history of conducting operations that maybe unauthorized and unethical.	"Based on the attitudes I have seen from AI companies and the way AI companies have already used data without consent, I'm unsure if they will respect robot.txt"
Profit	AI companies have monetary interests in scraping artists' work.	"Money before morals."
Perception	Artists perceive AI companies negatively (e.g., as greedy or unethical).	"AI companies are morally bankrupt."
Loophole	AI companies might find loopholes or workarounds to bypass robots.txt.	"They might start loopholes to get around it or something "
Legal enforcement	The need and lack of legislation or legal enforcement.	"Generative AI is built on top of copyright infringement-they can't be profitable without it, so they will argue against any thing that prevents them from scrapping. They have to be forced to respect it by law, we can't trust their good faith."
Voluntary nature	Robots.txt is a voluntary mechanism.	"At best it seems that robot.txt is just a warning sign, and will not entirely stop AI companies from deciding to scrape any particular content."
Misc	Additional reasons not covered above.	"I think, unfortunately, a lot of companies will not respect and will do it anyway."

Table 12: Codebook for why artists do not trust AI companies to respect robots.txt.