

Somesite I Used To Crawl:

Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers

Enze Liu* 

Elisa Luo* 

Shawn Shan 

Geoffrey M. Voelker 

Ben Y. Zhao  Stefan Savage 

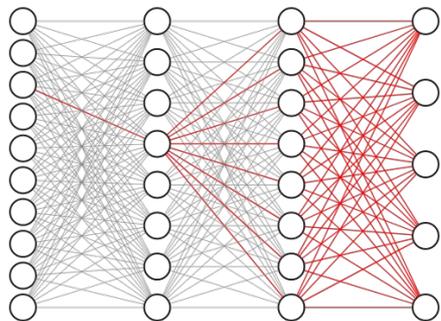
* The first two authors contributed equally and are listed alphabetically

 UC San Diego  University of Chicago

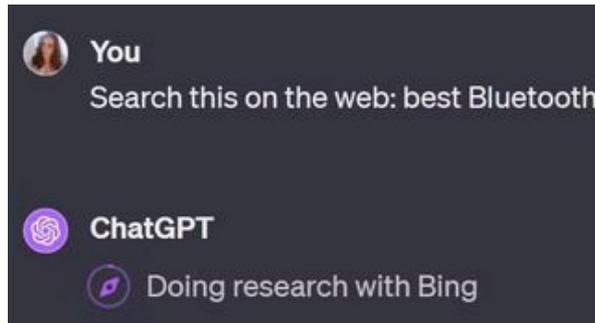
 Carnegie Mellon University

LLMs are Interacting w/ the Web Like Never Before

Training



Live Content Fetching



Automate Browser Tasks



Gen AI Creates Copyright Concerns

Copyright lawsuits

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Disney and Universal sue AI image company Midjourney for unlicensed use of Star Wars, The Simpsons and more

Companies crawl secretly

Perplexity AI Is Lying about Their User Agent

15th June 2024

Impacts on site performance

AI Crawlers Are Reportedly Draining Site Resources & Skewing Analytics

Gen AI Creates Copyright Concerns

Copyright law

How do *content creators* defend against *AI-related crawling*?

Analytics

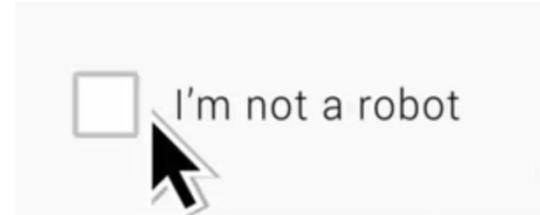
Content Protection Mechanisms

Robots.txt

(RFC 9309)

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

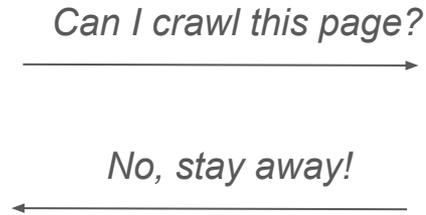
Active Blocking



One Possible Solution: Robots.txt

AI Data Crawler

Robots.txt File



One Possible Solution: Robots.txt

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

Robots.txt File

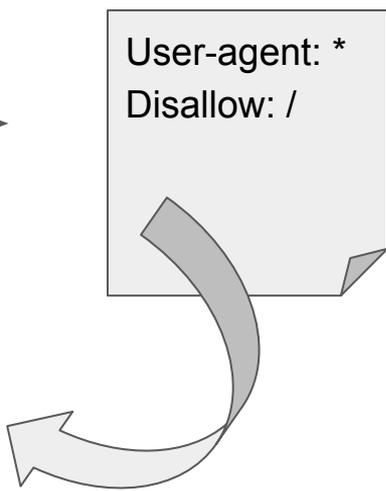
```
User-agent: *  
Disallow: /
```

One Possible Solution: Robots.txt

```
User-agent: Googlebot  
Allow: /  
  
User-agent: GPTBot  
User-agent: CCBot  
Disallow: /  
  
User-agent: *  
Disallow: /secret/
```

Robots.txt File

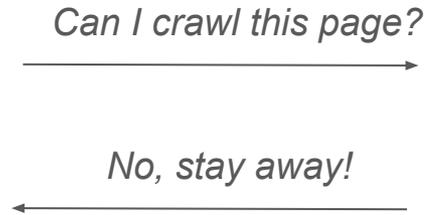
```
User-agent: *  
Disallow: /
```

A diagram illustrating the integration of a smaller robots.txt file into a larger one. A smaller, light gray box with a folded bottom-right corner contains the text 'User-agent: *' and 'Disallow: /'. A thick, curved arrow points from this smaller box towards the larger box on the left, indicating that the contents of the smaller file are being added to or merged with the larger file.

One Possible Solution: Robots.txt

AI Data Crawler

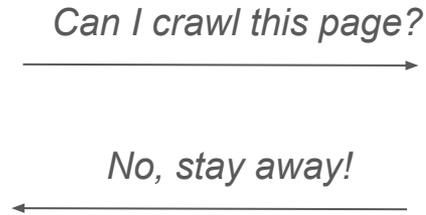
Robots.txt File



One Possible Solution: Robots.txt

AI Data Crawler

Robots.txt File



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



This Mechanism is Voluntary!

One Possible Solution: Robots.txt

AI Data Crawler

I'll crawl
anyways
~_(\ツ)_/\

Robots.txt File

Can I crawl this page?

No, stay away!

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



This Mechanism is Voluntary!

RQ1: Do content creators use robots.txt?

Well-resourced Domains



Artists



RQ1: Do content creators use robots.txt?

Well-resourced Domains



Population 🧑🧑: Stable Top 100k Sites from Sept 2022 to Oct 2024

Dataset 📁: Common Crawl (~monthly snapshots)

Methodology 📊: Check for restrictions on AI-related crawlers in robots.txt

RQ1: Do content creators use robots.txt?

Types of AI-Related Crawlers

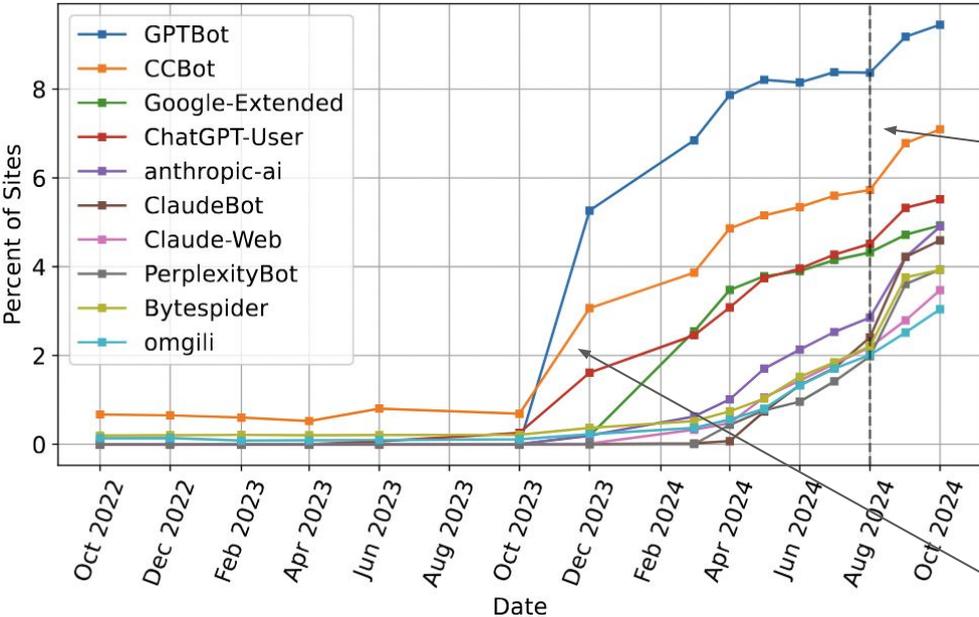
- 1) **AI data crawlers.** E.g., GPTBot
- 2) **AI assistant crawlers.** E.g., ChatGPT-User
- 3) **AI search crawlers.** E.g., OAI-SearchBot

Population 🐜🐜: Stable Top 100k Sites from Sept 2022 to Oct 2024

Dataset 📁: Common Crawl (~monthly snapshots)

Methodology 📊: Check for restrictions on AI-related crawlers in robots.txt

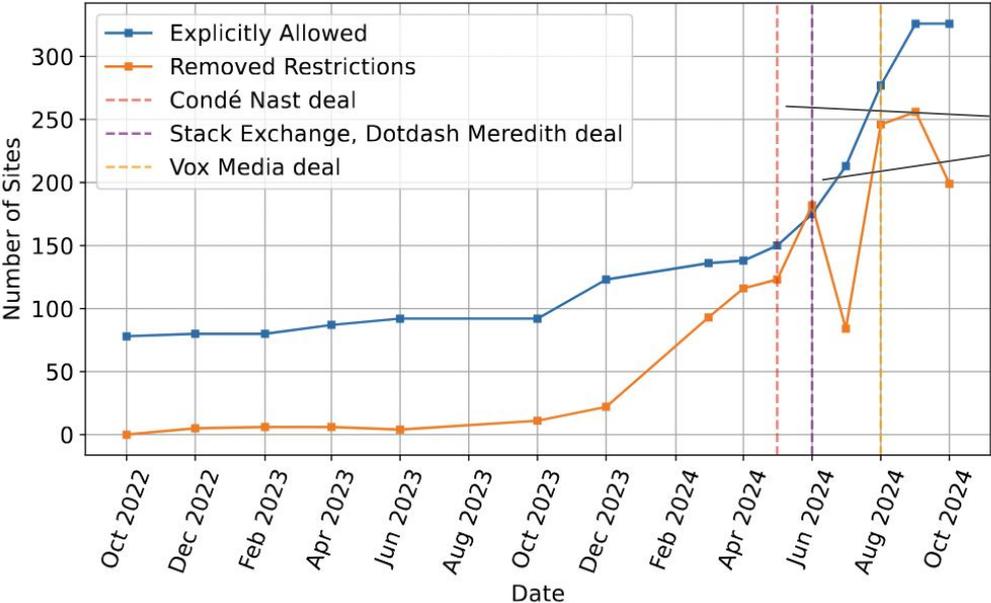
Increasing Drive to Protect Content



Sub-Measure 4.1. **Respect Robots.txt**

Release of GPTBot + ChatGPT-User

Recent Decrease in Restrictions



Possible Explanations:



Partnerships with AI Companies



Reverse Intent

RQ1: Do content creators use robots.txt?

Well-resourced Domains



~14% of top 5k Sites
~10% of top 100k Sites

Artists



RQ1: Do content creators use robots.txt?

Do artists want to stop AI-crawling?

- 96% (175) would use a tool that can block AI crawling

Have artists adopted robots.txt?

- 6 out of 182 utilized robots.txt

What challenges do they face?

- Awareness, Ability, Agency

Artists



RQ1.2: Do artists use robots.txt? Why or why not?

Awareness

60% Not Aware

Ability

“i dont know how to do it”

Agency

“Squarespace does not allow users to edit the robots.txt file.”

RQ1.2: Do artists use robots.txt? Why or why not?

Awareness

60% Not Aware

Ability

“i dont know how to do it”

Agency

“Squarespace does not allow users to edit the robots.txt file.”



SQUARESPACE

Option to “Block AI”
(Only 17% enable)



Edit robots.txt
(paid ver. only)



Can't edit robots.txt



Can't edit robots.txt



Can't edit robots.txt



Can't edit robots.txt



ARTSTATION

Can't edit robots.txt

RQ2: Do AI-crawlers respect robots.txt?

Test on a website we control with robots.txt

Passive testing *Wait*

for crawler to visit

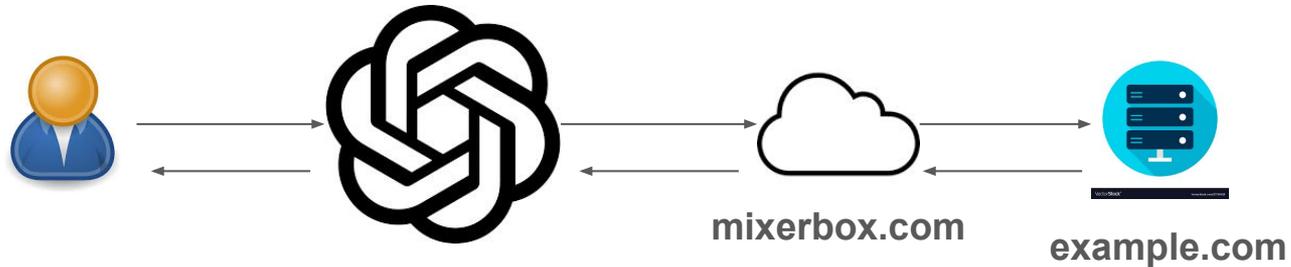
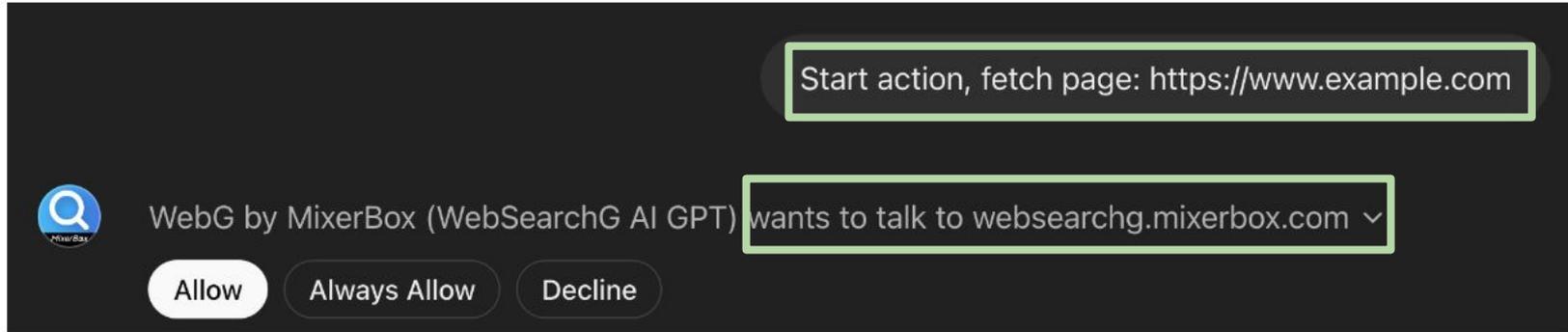
- AI Data Crawler
(e.g., GPTBot)
- AI Search Crawler
(e.g., OAI-SearchBot)

Active testing

Trigger crawler visit

- AI Assistant Crawler (e.g., ChatGPT-User)
- **Third-party AI Assistant Crawler**

Background: Third-party AI Assistant Crawler



RQ2: Do AI-crawlers respect robots.txt?

- **Results — Passive Testing**

- Attracted 7 different crawlers



Amazonbot



OpenAI

GPTBot



COMMON
CRAWL

CCBot



Claude

ClaudeBot



AppleBot



Meta-ExternalAgent



TikTok

Bytespider



RQ2: Do AI-crawlers respect robots.txt?

- Results — **Active Testing**
 - ChatGPT-User 😊
 - 21 Third-party AI Assistant Crawlers
 - **Only 2** fetched robots.txt

RQ2: Do AI-crawlers respect robots.txt?

amazon

Amazonbot



OpenAI

GPTBot



COMMON
CRAWL

CCBot



Claude

ClaudeBot



AppleBot



Meta

Meta-ExternalAgent



TikTok

Bytespider



Third-party AI Assistant
Crawlers

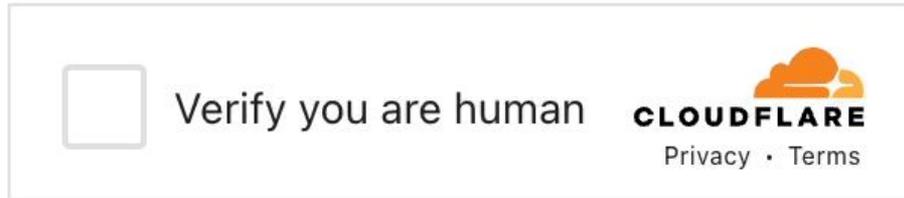


RQ2: Do AI-crawlers respect robots.txt?

a

Do we have tools with stronger protections?

Active Blocking — An Alternative to Robots.txt



Active Blocking — An Alternative to Robots.txt

New AI Scrapers and Crawlers

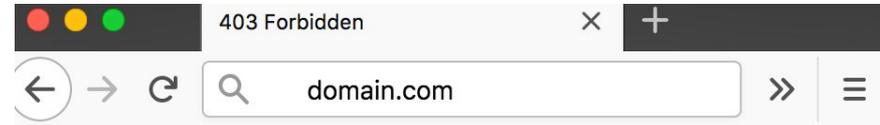
Block bots from scraping your content for AI applications like model training.

Verify you are human

CLOUDFLARE
Privacy · Terms

I'm not a robot


reCAPTCHA
Privacy - Terms



Forbidden

You don't have permission to access / on this server.

RQ3: Do content creators utilize active blocking?



RQ3: Do content creators utilize active blocking?

- **14%** of top 10k sites use active blocking against AI-related bots
- **Only 2%** of these sites have corresponding restrictions in **robots.txt**



Using
robots.txt



Using
active
blocking



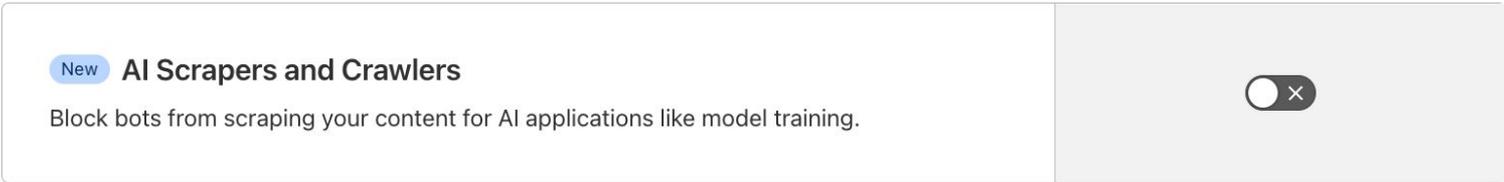
Third-Party Active Blocking: **CLOUDFLARE**

Operation

Who is considered an “AI Scraper [or] Crawler”?

Adoption

How many sites have enabled this option?



- 2,018 (20%) of Top 10k sites use Cloudflare



Third-Party Active Blocking: CLOUDFLARE®

Operation

Who is considered an “AI Scraper [or] Crawler”?

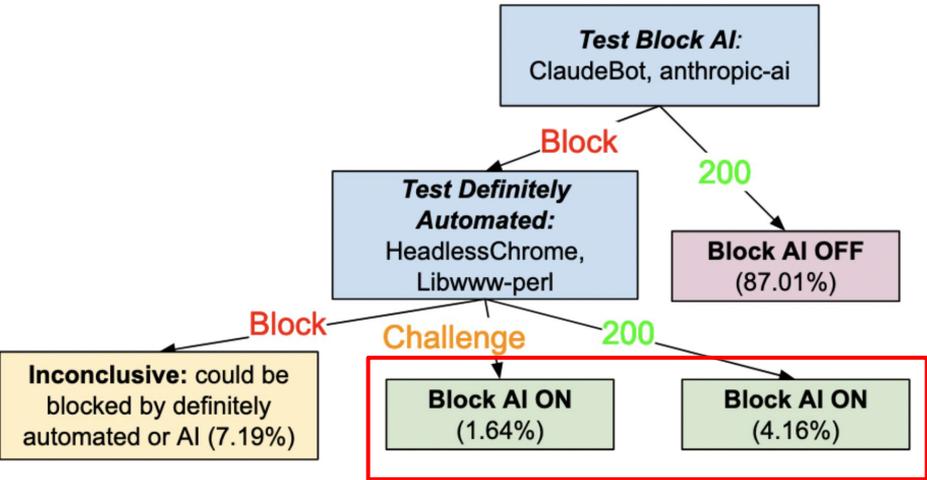
Amazonbot	Diffbot/
AwarioRssBot	GPTBot/
AwarioSmartBot	magpie-crawler
Bytespider	MeltwaterNews
CCBot/	omgili/
ChatGPT-User	PerplexityBot
Claude-Web	PiplBot
ClaudeBot	YouBot
cohere-ai	

17 AI-related User-Agents*

* as of Oct 2024³²



Third-Party Active Blocking: CLOUDFLARE®



Adoption
How many sites have enabled this option?

At least **5.7%** of sites using Cloudflare enable “Block AI”

Where AI Crawler Restriction Mechanisms



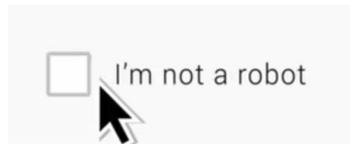
Robots.txt (~11%_{Top10k})

- Voluntary
- Syntactic ambiguity
- What is a “robot”?

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

Active Blocking (~14%_{Top10k})

- Multi-purpose crawlers (e.g. Applebot)
- Impacts on SEO



General

- Crawlers must self-identify
- No Take-Backs
- Legal uncertainty

Somesite I Used to Crawl – Summary

AI-related crawling generates **copyright** concerns



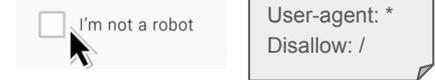
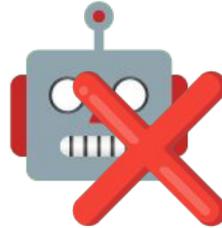
Need for mechanisms to **control** AI-related crawling



Use robots.txt and active blocking as **ad-hoc solution**

Disney and Universal sue AI image company Midjourney for unlicensed use of Star Wars, The Simpsons and more

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work



Robots.txt and active blocking have their own **limitations**

Somesite I Used To Crawl:

Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers

Enze Liu*  Elisa Luo*  Shawn Shan  Geoffrey M. Voelker 

Ben Y. Zhao  Stefan Savage 

* The first two authors contributed equally and are listed alphabetically

 UC San Diego  University of Chicago

 Carnegie Mellon University

To Appear in the Proceedings of the Internet Measurement Conference 2025

Questions?

e4luo@ucsd.edu



Thank You!

[elisa-luo.github.io](https://github.com/elisa-luo)

<https://arxiv.org/pdf/2411.15091>