A Klick Health POV

GPTBot

—

To Block or Not to Block

Sharon Virtue
Vice President, SEO
Klick Health

Shu Ito
Vice President, Science & Regulatory
Klick Health

GPTBOT: TO BLOCK OR NOT TO BLOCK

Currently, Klick recommends not blocking GPTBot on life sciences promotional websites as the inclusion of accurate, up-to-date, and truthful information as training data is imperative, especially where healthcare is concerned, to reduce biases in how future content is generated. Klick will continue to monitor, evaluate, and update this POV as other AI bots provide this option and we start to see the outputs of how this crawled information is used and presented to consumers.

As a logical evolution of ChatGPT, companies are now being given the choice to allow the popular AI platform to scrape their website content for training purposes. Some prominent websites are already reacting to this option by proactively disallowing the web crawler from having access to their data. Does this make sense for life sciences promotional websites? To help life sciences marketers answer the question, we have created this POV, which explores the technical options now available, and ultimately recommends not to block GPTBot on life sciences promotional websites.

BACKGROUND

OpenAI's ChatGPT is a type of Large Language Model (LLM). These systems work by trying to predict "what comes next" in a pattern of text. They are not search engines. They do not retrieve information they have seen in the past and present a copy of it like a search engine does. They are more akin to a contestant on a quiz show: constructing an answer based on their latent memory of what they have seen before.

On August 7, OpenAI released its web crawler called GPTBot, along with information on how brands can prevent it from scraping website content. Why is this important? The addition of crawling ability will provide ChatGPT with a much larger, robust dataset on which to train; this dataset includes life sciences websites.

When prompted "Can you crawl the web?" ChatGPT currently responds:

> **"As an AI developed by OpenAI, I don't have the capability to crawl the web in the way a search-engine bot like Googlebot or Bingbot does. My main function is to generate human-like text based on the prompts given to me. I can provide information and answer questions to the best of my knowledge and training, but I don't actively fetch or retrieve information from the internet in real time."**

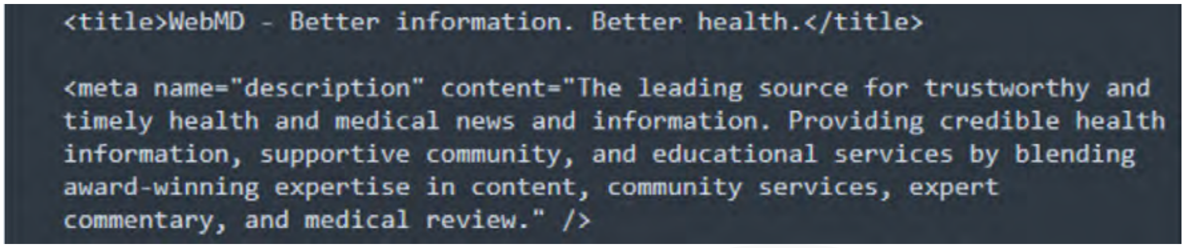
With the release of GPTBot, life sciences brands need to be aware of the options available to block GPTBot's ability to crawl the content of their sites and the associated considerations. Doing nothing means allowing GPTBot access to your data for training purposes.

UNDERSTANDING WEB CRAWLERS

Web crawlers (also known as spiders, robots, and bots) are automated programs, most commonly used by search engines like Google, Bing, and Baidu. They all crawl the web by following links and scraping content in order to enhance their indexes and serve up highly relevant organic search results based on keyword queries. Website owners do not have to ask for their websites to be crawled, but they can help speed up the indexing process in most cases by including a list of website pages in sitemap.xml. The frequency at which the crawler returns to a site will depend on many factors, including how often the content is refreshed and its perceived quality.

For life sciences websites, SEO teams create page-level metadata as “spider food” for a brand’s organic listings. The metadata, which includes page title tags, meta descriptions, and URLs, goes through the MLR submission process and, once approved, is implemented as code on each page of the website. It should be noted that a search engine can tweak any approved SEO metadata to present in the search-engine results page (SERP)—this is out of a brand’s control. The best defense to prevent this from happening is to create metadata that closely aligns with the page’s content.

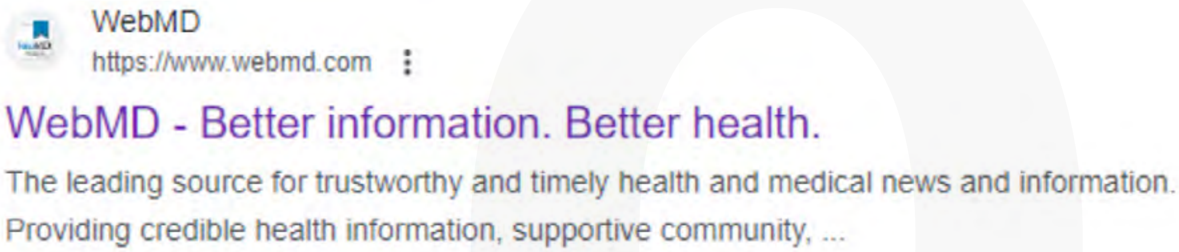
Example of metadata within a website page’s code:



```
<title>WebMD - Better information. Better health.</title>

<meta name="description" content="The leading source for trustworthy and
timely health and medical news and information. Providing credible health
information, supportive community, and educational services by blending
award-winning expertise in content, community services, expert
commentary, and medical review." />
```

Example of a resulting organic listing on a Google SERP:



WebMD
https://www.webmd.com

WebMD - Better information. Better health.

The leading source for trustworthy and timely health and medical news and information. Providing credible health information, supportive community, ...

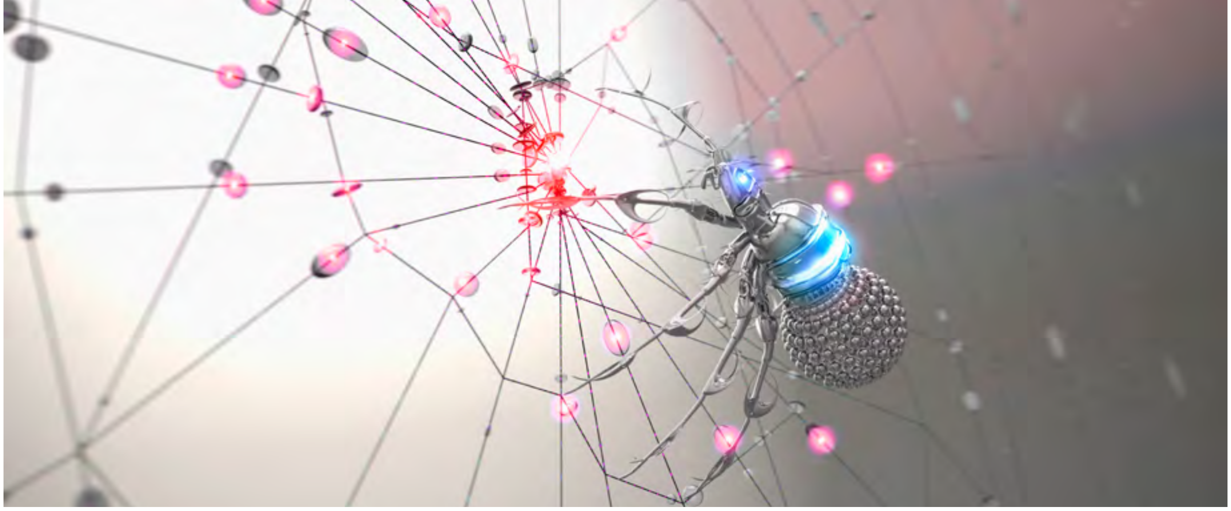
ENTER THE NEW WEB CRAWLER: THE AI BOT

Much like ChatGPT, GPTBot made its debut with both splash and controversy, in part because its release was just quietly slipped into OpenAI documentation rather than being formally announced:

> **"Web pages crawled with the GPTBot user agent may potentially be used to improve future models and are filtered to remove sources that require paywall access, are known to gather personally identifiable information (PII), or have text that violates our policies. Allowing GPTBot to access your site can help AI models become more accurate and improve their general capabilities and safety. Below, we also share how to disallow GPTBot from accessing your site."**

Unlike a Googlebot, it is currently unknown where and how scraped content might appear as a result of AI prompts as well as whether the content provider will be referenced in any way. To allay brand concerns, OpenAI is providing the option to disallow crawling of website content through the use of the robots.txt file.

The robots.txt file is a standard way of communicating to any web crawler which parts of a website they should not crawl or index. Not all bots (certainly not malicious ones) will obey the directives within a robots.txt file, but the major global search engines do—as does GPTBot.



OPTIONS FOR GPTBOT CRAWLING WEBSITE CONTENT

1. To disallow GPTBot from crawling an entire domain, add this code snippet to the robots.txt file:

```
User-agent: GPTBot
Disallow: /
```

As live examples at time of publication, both Healthline and WebMD have disallowed GPTBot in their robots.txt files. This choice serves to protect data they wish to monetize.

2. To disallow GPTBot only from specific sections/pages of a website, think of GPTBot as any other web crawler and keep it out of folders that aren't intended to be public, have no content value, and/or could trap the bot in an endless loop.

Example:

```
User-agent: *
Disallow: /admin/
Disallow: /tmp/
Disallow: /test/
```

3. Alternatively, to allow GPTBot to crawl the content of a domain in its entirety, the directive would be:

```
User-agent: *
Allow: /
```

IMPORTANT TO NOTE:

- The above allows ALL user-agents to crawl website content, not just GPTBot
- Once data is part of GPT's training, it currently does not appear that website content can be removed. This is unlike the search engines, which all have the ability to request removal of website pages from their indexes

POTENTIAL REGULATORY CONSIDERATIONS

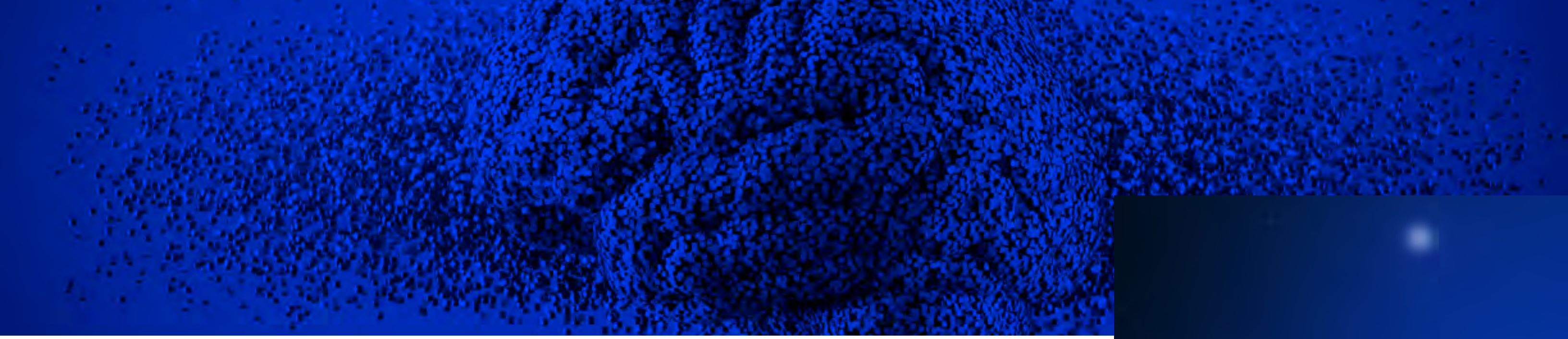
There have been some immediate reactions to the news about being able to prevent GPTBot from crawling websites, based on fears around how the information will be used and presented. However, given what we know about LLMs, we believe that it is unlikely that ChatGPT, when trained using information on a life sciences website, will output significant sections of the website's content. It may, however, learn small phrases or facts from the content.

Nevertheless, from a regulatory perspective, the FDA guidelines state that product claims must be fairly balanced with safety information, presented with equal prominence to the claim. These unknowns may prompt regulatory teams to adopt a stance to disallow access to the life sciences websites.

Despite these unknowns, life sciences companies may not be responsible for such content generated by an AI such as ChatGPT. The FDA clearly defines information created by third parties as user-generated content (see https://www.fda.gov/regulatory-information/search-fda-guidance-documents/internetsocial-media-platforms-correcting-independent-third-party-misinformation-about-prescription). In addition, they have stated:

> **"Firms are generally not responsible for third-party UGC about their products when the UGC is truly independent of the firm (e.g., is not produced by, or on behalf of, or prompted by the firm in any particular) regardless of whether the firm owns or operates the platform on which the communication appears."**

Content generated by ChatGPT that is not produced by, on behalf of, or prompted by a life sciences company (i.e., truly independent), by definition would fall under third-party UGC, and would not be the responsibility of the company. Although the usage of crawled data is currently not known, this should not constitute the sole reason for disallowing access to a life sciences company's website.



COMMITMENT TO PROVIDING ACCURATE INFORMATION

A number of companies have reportedly blocked GPTBot, and it is reasonable to ask if life sciences companies should follow suit. The companies blocking the crawler have something in common: they seek to monetize the data on their websites, often as content around which ad impressions can be sold as well as through subscriptions and data analytics.

This motivation contrasts that of public-facing life sciences websites whose objective is to provide accurate information. The provision of accurate information is supported by a rigorous review process that ensures all consumers, caregivers, and healthcare professionals have access to scientifically valid, up-to-date, and as FDA describes, truthful and non-misleading information.

Eliminating sources of accurate and compliant information by disallowing GPTBot access may bias AIs, such as ChatGPT, to create content based on less rigorously reviewed information, and in some cases, misinformation. Just consider how in 2018, an AI developed by IBM was providing unsafe and incorrect cancer treatments based on a small number of hypothetical patients with cancer that were used to train the AI (https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/). As early as 2016, Microsoft developed a chatbot designed to interact with users on Twitter, but quickly canceled the experiment after the chatbot learned to produce inflammatory comments using unfiltered data from its interactions on the platform.

KEY TAKEAWAYS

This is only the beginning. There are many other AI platforms beyond OpenAI/GPT that are (or will be) looking for publicly accessible content to train their models. Disallowing GPTBot does not guarantee that a website's content will not be accessed, or that it hasn't already been scraped by an AI bot. ChatGPT itself was trained by scraping the web up to September 2021—a process that included life sciences websites.

AIs are trained by reading everything and then weighing the totality of the information to learn. Large authoritative health-information sites, such as Healthline and WebMD, have massive and detailed amounts of data for AIs to train on. We can expect these websites to be highly weighted by AIs.

Finally, although robots.txt files are regularly utilized by IT teams and the SEO industry to control which pages a spider is allowed to visit and index, it is not a foolproof mechanism for a few reasons:

- Complex directives (disallowing certain crawlers from specific areas of a domain) can unintentionally keep crawlers out of sections of a website you would want to be crawled/indexed or invite them to visit private areas (e.g., admin)
- Robots.txt files are notorious for being accidentally overwritten to a previous version with website updates if appropriate deployment best practices are not adopted
- Robots.txt is a voluntary standard, and some web crawlers will ignore it

Be certain to have expert advice on how to format a robots.txt file if choosing any option to disallow, and regularly monitor the robots.txt file for changes.

RESOURCES

- OpenAI GPTBot Documentation
- Google Search Central Documentation Googlebot
- OpenAI Introduced GPTBot Web Crawler to Index Websites
- Sites Scramble to Block ChatGPT Web Crawler after Instructions Emerge
- Understanding OpenAI's GPTBot & Robots.txt Setup
- The New York Times Blocks OpenAI's Web Crawler

UPDATES

SEPTEMBER 22, 2023

[Announcing new options for webmasters to control usage of their content in Bing Chat]

Bing has chosen a different route than ChatGPT and Google, providing web publishers with control through the meta robots tag instead of the robots.txt file. Although page level tagging is potentially a heavier lift for webmasters, it does allow for a more granular level of control.

SEPTEMBER 28, 2023

[AI: An update on web publisher controls]

Google has announced Google-Extended, its own mechanism for web publishers to control crawling of Bard and Vertex AI generative APIs. Similar to GPTBot, instructions for disallowing the crawling of content can be stipulated in the robots.txt file.

A FINAL WORD:

Klick currently recommends not blocking AI creators from crawling content on life sciences promotional websites. Now that ChatGPT, Google and Bing have all come out with a position we can expect other AIs to follow suit in the future

klick HEALTH

We welcome your questions
and feedback. Please contact:

Michael Chambers
SVP, Opportunity Creation
mchambers@klick.com