



NOTE

Content Signals in robots.txt: what the proposal says and what it can enforce

The Content Signals Policy adds a line to robots.txt that expresses preferences by purpose rather than by crawler: three named signals (search, ai-input, ai-train) each carry a yes or no, stating whether content may be used for search, model input, or training.

KIND	Note
DATE	2026
SUBJECTS	Crawlers & access; Conventions & policy
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

The Content Signals Policy adds a line to robots.txt that expresses preferences by purpose rather than by crawler: three named signals (search, ai-input, ai-train) each carry a yes or no, stating whether content may be used for search, model input, or training. This is a different axis from the per-token crawler groups the taxonomy (BA-C-6) classifies: one names purposes and reaches every crawler, the other names crawlers and is silent on purpose. This note reads the published policy, states who authored it and its relation to the IETF effort it points to, and places it on robots.txt's enforcement boundary: a request and, in the policy's framing, a reservation of rights, not an access control. A re-analysis of this publication's census corpus adds a first field measurement: a Content-Signal line on 3.0% of parsed domains, most carrying a CDN's managed default that also deploys a fourth key beyond the announced three.

CONTENTS

1	What the proposal adds	3
2	Who published it, and its relation to a standard	3
3	A different axis from the per-token model	3
4	What it can and cannot enforce	4
5	Adoption in the wild: a first measurement	4
6	Limitations	5

A robots.txt file, as the companion note on [robots.txt](#) and [AI crawlers](#) describes it, addresses named crawlers: a rule disallows GPTBot, or OAI-SearchBot, or CCBot, each a token standing for a piece of software with a documented job. The Content Signals Policy, published by Cloudflare in September 2025, adds a second kind of statement to the same file — one that names a *purpose* instead of a crawler, and applies to every crawler at once. It is a small syntactic addition with a different conceptual shape — worth stating precisely, because the two constructs answer different questions and are easily conflated.

1 What the proposal adds

The policy defines a single new line for robots.txt. Its form is a comma-separated list of named signals, each set to yes or no — the policy’s own example reads `Content-Signal: search=yes, ai-train=no` [1]. Three signals are defined, each given a meaning in explanatory text the policy embeds as robots.txt comments above the line. In the policy’s words, `search` covers “building a search index and providing search results (e.g., returning hyperlinks and short excerpts from your website’s contents),” with the added qualification that “Search does not include providing AI-generated search summaries”; `ai-input` covers “inputting content into one or more AI models (e.g., retrieval augmented generation, grounding, or other real-time taking of content for generative AI search answers)”; and `ai-train` covers “training or fine-tuning AI models” [1].

The three yes/no/absent states are defined explicitly. Per the embedded policy text, “If a content-signal = yes, you may collect content for the corresponding use” and “If a content-signal = no, you may not”; and where a signal is omitted, “the website operator neither grants nor restricts permission via content signal with respect to the corresponding use” [1]. An omitted signal is therefore a documented non-statement, not an implied default — the same three-valued discipline (yes, no, unstated) the crawler taxonomy requires of a registry distinguishing a declined control from one never offered.

2 Who published it, and its relation to a standard

The policy is Cloudflare’s, authored by Will Allen and published on the company’s blog on September 24, 2025 [1]. It is a vendor policy, not an adopted standard — a distinction that bears on how much weight the vocabulary carries. Cloudflare applies it by default for customers using its managed robots.txt feature, with `search=yes` and `ai-train=no` set and `ai-input` left unset because, the company writes, it did not want to assume that preference on customers’ behalf [1]. The same post points to the venue where standardization, if it happens, would occur: the IETF’s AI Preferences working group, chartered to “standardize building blocks that allow for the expression of preferences about how content is collected and processed for Artificial Intelligence (AI) model development, deployment, and use,” including a vocabulary and the means of “attaching or associating those preferences with content,” explicitly naming the Robots Exclusion Protocol among those means [2]. Content Signals should thus be read as one vendor’s concrete proposal in a space the IETF is working to standardize — related to that effort, but not itself its output.

3 A different axis from the per-token model

The crawler taxonomy (BA-C-6) sorts crawlers into functional classes — training, retrieval, user-fetch, agent, search — and notes that robots.txt controls them one token at a time: a rule names GPTBot to reach training, OAI-SearchBot to reach retrieval, and so on. That is audience-by-crawler. Content Signals is audience-by-purpose. It says nothing about which crawler is asking, and instead states, for any crawler, whether content may be used for a search index, as model input, or for training. The two are orthogonal, and neither reduces to the other. A per-token rule cannot express “no training, by anyone” without

enumerating every training crawler, present and future; a content signal cannot single out one crawler, because it does not name crawlers at all. The purposes map loosely onto the taxonomy's classes — ai-train onto training, ai-input onto the retrieval and user-fetch classes, search onto the search class — but the mapping is many-to-many, and the constructs stay distinct: one names crawlers and is silent on purpose, the other names purposes and is silent on crawlers.

4 What it can and cannot enforce

Whatever axis it uses, a content signal sits inside robots.txt and inherits robots.txt's limits. The governing specification states that its rules “are not a form of access authorization” [3]; a line in the file is a request a crawler is asked to honor, not a barrier the file enforces, and an unrecognized field is one a conforming parser may simply ignore. Cloudflare is explicit that its policy does not change this: content signals, the company writes, “express preferences; they are not technical countermeasures against scraping” [1]. What the policy adds on top of the request is a legal characterization, not a technical one: its text asserts that restrictions expressed via content signals are “EXPRESS RESERVATIONS OF RIGHTS UNDER ARTICLE 4 OF THE EUROPEAN UNION DIRECTIVE 2019/790 ON COPYRIGHT AND RELATED RIGHTS IN THE DIGITAL SINGLE MARKET” [1]. Whether that characterization has the legal effect it claims is a question of copyright law in a specific jurisdiction, not a property of the file, and nothing in this note is legal advice. The technical fact is unchanged from the robots note's conclusion: compliance is a decision made by whoever operates a crawler.

5 Adoption in the wild: a first measurement

A first field measurement of how the proposal appears in practice comes from a re-analysis of the BA-D-2026-01 raw corpus — a 2026-07-10 run over the robots.txt files behind this publication's crawler-access dataset, performed as a raw-text line scan rather than a parser pass. Among the 1,381 domains whose robots.txt parsed cleanly, a Content-Signal line appears on 42 — 3.0% — across universities (10), news (16), e-commerce (8), and government (8). Cloudflare separately reports that “over 3.8 million domains” use its managed robots.txt feature to express a training opt-out [1]; the 3.0% is a measured rate in a defined sample, not that vendor service-base count, and the two are different quantities.

The composition of the 42 is the more telling result. Thirty-three of them — 79% — carry Cloudflare's managed-content signature, and all thirty-three set the identical value search=yes, ai-train=no, use=reference. The most common content-signal string in the wild is thus not an operator's considered expression of the policy; it is one CDN's managed default, applied uniformly. The nine hand-authored carriers, by contrast, use only the three documented keys and vary among themselves — values include search=yes, ai-input=yes, ai-train=no, one setting ai-input=no, and in one case ai-train=yes. And the signal is layered on blocking rather than replacing it: 40 of the 42 also root-block at least one AI crawler token in the same file.

That dominant value carries a fourth field the policy this note opened on does not define, and the gap is itself the finding. The Content Signals Policy as announced in September 2025 defines three keys — search, ai-input, ai-train — and no more [1]. Cloudflare's deployed managed robots.txt documents a fourth in its own embedded comment block: the corpus's archived bytes carry the self-definition use: how AI systems may consume the content (immediate, reference, or full) above the directive, and all 33 managed carriers set use=reference. That key is a separate, optional Cloudflare extension, “content-use,” whose documentation defines reference as “Index, excerpt, and link back” and which Cloudflare “adds ... to the managed content, in line with the existing default of search=yes, ai-train=no” [4]. The

version difference is thus concrete: the announced policy carries three keys, the most-deployed managed text in this corpus carries four, and the extra key is a CDN default rather than anything operators authored — which is also why the hand-authored files above stay within the announced three. The corpus shows the space is unsettled in other ways too: the same scan found an `x-rs1` field on 7 news domains, a content-usage field on 3 e-commerce domains, and a license field on 3 news domains — small counts, but more than one vocabulary in play at once.

6 Limitations

This note reads a single vendor policy and one working-group charter as published and archived on 2026-07-10, alongside the robots specification. Both the policy text and the standardization effort are moving: the vocabulary may change, and the IETF work may supersede or absorb it. The wild-adoption figures come from one corpus (BA-D-2026-01, four sectors, the 1,381 domains that parsed), read by a raw-text line scan that counts the presence and literal value of a Content-Signal line — not by a conformance parser, and not as evidence that any crawler acted on the line. The relationship drawn here between the policy's purpose-signals and the taxonomy's crawler-classes is conceptual, not a claim that any crawler treats the two consistently. And whether crawlers honor the signal, and whether the copyright reservation holds, are separate questions this note does not answer.

REFERENCES

1. Cloudflare (W. Allen). *Giving users choice with Cloudflare's new Content Signals Policy* (2025). <https://blog.cloudflare.com/content-signals-policy/> Accessed 2026-07-10. [archived]
2. IETF AI Preferences (aipref) Working Group. *AI Preferences (aipref) — About / Charter* (2025). <https://datatracker.ietf.org/wg/aipref/about/> Accessed 2026-07-10. [archived]
3. M. Koster, G. Illyes, H. Zeller, and L. Sassman, IETF. *RFC 9309: Robots Exclusion Protocol* (2022). <https://www.rfc-editor.org/rfc/rfc9309.html> Accessed 2026-07-08. [archived]
4. Cloudflare (Cloudflare Docs). *Managed robots.txt — content signals and the content-use extension* (2026). <https://developers.cloudflare.com/bots/additional-configurations/managed-robots-txt/> Accessed 2026-07-10. [archived]

HOW TO CITE

Barkhausen AI (2026). Content Signals in robots.txt: what the proposal says and what it can enforce.

<https://barkhausen.ai/notes/content-signal-analysis/>

Published under the Creative Commons Attribution 4.0 International (CC-BY-4.0).

