TORQUE

GEO ELITE
Nuances of Bot Management for Search

To Block or not to block is… no longer the only question.

Chris Green

01

The model has broken

Why block-or-allow stopped being enough

“Block anything you don't like or recognise.

Or only allow what you do.”

Bot management for SEO, any time before about 2024.

I'm not advocating for either.
It was simply all we needed.

That model does not describe the web we work with now

Arriving at your site today:

- **Search crawlers**
- **AI retrieval crawlers**
- **Training crawlers**
- **User-initiated fetches**
- **Agents and agentic browsing**

The Block/Allow model isn't fit for purpose.

If you've worked with CDN-level bot mitigation, it has been unfit for a while.

02

“AI bot” is not a category

Why that term can cause more issues than it solves

“Are we blocking AI Bots?”

We know what this roughly means.

We just can't give a single yes/no answer.

Different bots from the same provider have completely different purposes —
and providers haven't aligned on what any of it looks like.

OpenAI, to their credit, make this clear

01	GPTBot	Model development. Looking for content to train on.
02	OAI-SearchBot	Building the search index behind ChatGPT's search.
03	ChatGPT-User	User-initiated retrieval. Fired when ChatGPT fetches a page for someone.

Very specific, very different, relatively easy to distinguish parts of the cycle.

Nobody else is quite that clean

Google splits differently:

- Googlebot
- Google-Extended
- and a various few other bits

No simple delineation of the kind OpenAI offer.

Then ask the same of:

- Perplexity
- Bing
- Anthropic
- Grok
- Muse

There is no real standard...

It is not as simple as it could or should be. But I guess that is the web for you.

It’s not impossible to decipher

Provider	Training	Search index	User fetch	Agent
OpenAI	GPTBot	OAI-SearchBot	ChatGPT-User	—
Anthropic	ClaudeBot	Claude-SearchBot	Claude-User	—
Perplexity	PerplexityBot	PerplexityBot same bot	Perplexity-User	—
Google	Google-Extended	Googlebot	—	—
Apple	Applebot-Extended	Applebot	—	—
Meta	Meta-ExternalAgent	Meta-ExternalAgent same bot	Meta-ExternalFetcher	—
Common Crawl	CCBot	—	—	—

Dashed = a robots.txt control token with no HTTP user agent of its own.

Nobody has standardised the agent column.

Even “do we block ChatGPT?” is wrong on several levels

It's a good way to start the conversation.

It is a terrible way to finish it.

It requires everybody involved to hold a view - and to be educated on what that view actually means, and what the implications are.

And this is the simple part.

What about the bots that aren't run by AI providers at all?

The bot nobody recognises

CCBot

Common Crawl's crawler. The data source it builds feeds AI training and infrastructure across the industry.

What happens if we block it?

Is anyone on your team even aware it's a useful bot - or did someone see it hitting the site regularly and block it because they didn't know who it was?

It may take months (or longer) to establish whether this means your fresh content is no longer available in model training data though.

I can assure you that happens.

I've seen it in the data – more than 3% of pages analysed from the HTTP Archive data

https://www.linkedin.com/posts/chrisgreenseo_most-people-arent-giving-common-crawl-the-share-7488481074797252609-bUhO/

03

robots.txt is not an effective controller

Even if bots are well behaved… which they're often not!

How comfortable are we to rely on robots.txt file to enforce business decisions like these?

The questions behind every line in it:

- Do we want our content discoverable?
- Do we want it usable in AI answers?
- Do we want it available for AI training?
- Do we want agents accessing our site on behalf of users?

Blocking one of these may be desirable and blocking another may be fundamentally damaging. Or worse, it might not be blocking what is needed at all!

Not all of these decisions are “are we found in search results or not”.

Some of it is commercial, branding or legal.

Suddenly robots.txt has active stakeholders

SEO	**IT/Security**	**Engineering**
Legal	**Commercial**	**Analytics**

These choices increasingly involve people who have never opened the file or are aware of its existence. Now we're inviting them into a conversation that was never as simple as block or allow.

A lot of businesses aren't prepared for that.

Robots Path

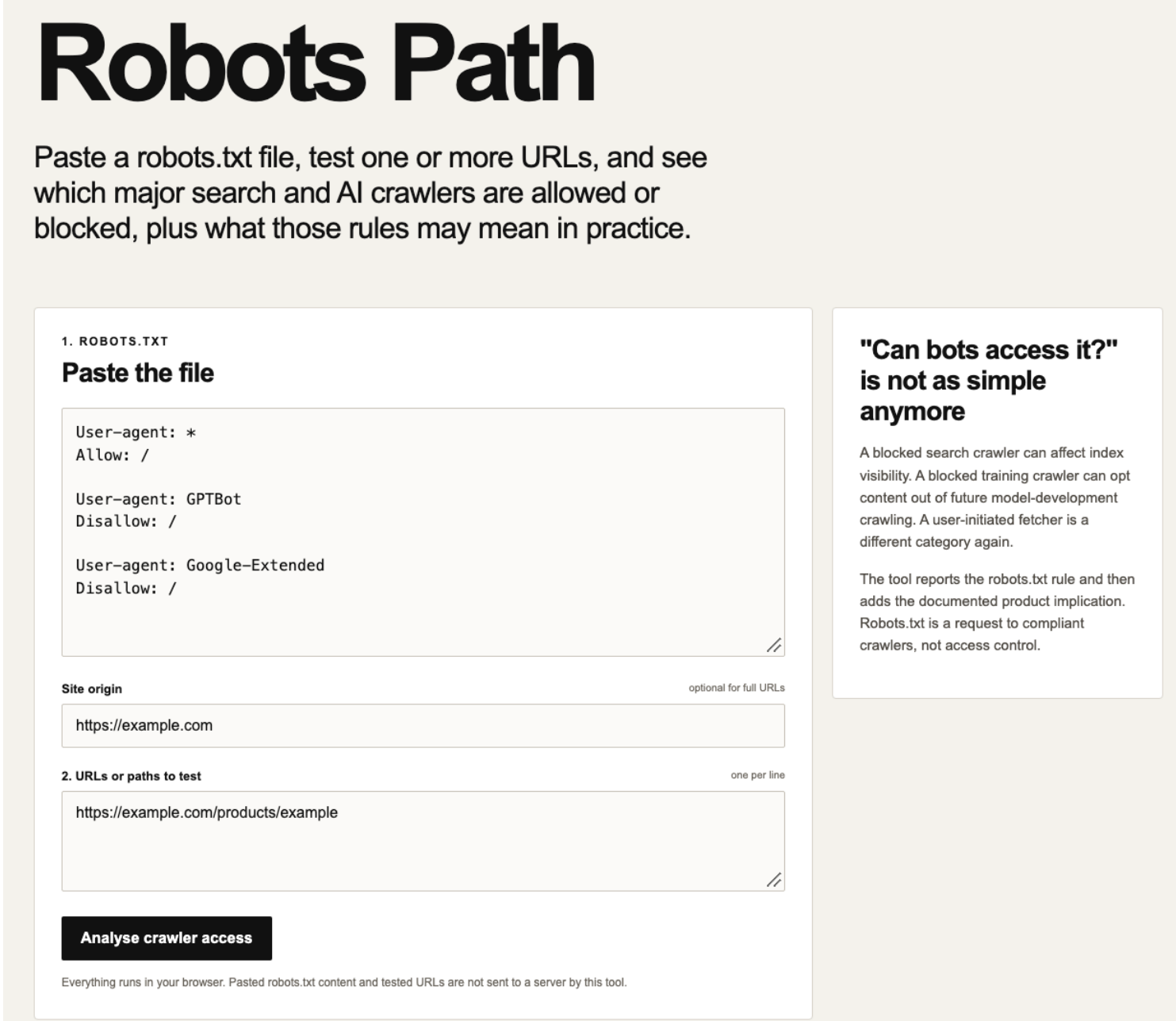
1 Paste your robots.txt into the tool

2 Set your domain

3 Paste in URLs from your site

4 See which parts of the AI search ecosystem you're blocking versus allowing

Use this small, free tool to help understand what your robots.txt is blocking



You can read a robots.txt file, make assumptions from what it contains, and still be completely wrong.

Without understanding user-agent taxonomies, your own site structure, and what the rules actually match - you have may gain false confidence.

And every rule in it is voluntary

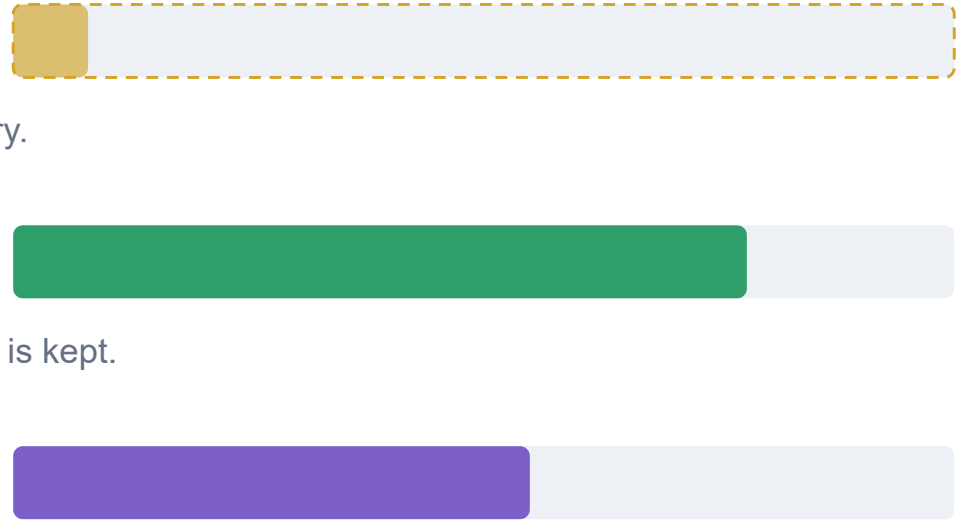
We call them rules, that is how we want them to be treated. But nobody crawling the web has to respect them.

- Some AI companies have been open about whether they comply
- Some have changed their view over time
- Some have actively misled historically or do now whenever they visit your site (Muse/Grok…)

It’s not to say we shouldn’t use a robots.txt… Just don't let it be the only place you try and manage bots

Three layers. Very different teeth.

Likely enforcement success - just illustrative purposes(!)



robots.txt

Advisory. Compliance is entirely voluntary.

CDN / WAF

Actual enforcement. Where the decision is kept.

Challenges & heuristics

Probabilistic. Catches some, frustrates others (potentially).

The layer SEOs are stuck with, is actually the least effective.

That is the whole problem with treating it as the control.

The control layer is further down

Where enforcement actually happens:

- Your CDN
- Web application firewall
- Any bot management layer in front of the site

Cloudflare, Akamai, Fastly built much of their business at protecting sites at this level.

What they give you:

- Bot rule definitions
- Validation of real bots versus spoofed ones
- A lot of heavy lifting out of the box

But their definition of a good bot may not be yours - and mistakes here have had devastating knock-on effects.

You can't have your cake and eat it

Plenty of businesses want to appear in search and be retrieved in AI answers… and simultaneously not be trained on.

Cloudflare make it really (too) easy for “disallow AI” to become a standard setup.

For a lot of websites that could be more damaging than not. If you aren't an informed user of these tools, you can block things you were never even aware of.

Publishers are most exposed.

If you rely on search traffic you need Google. Inhibit it from training or retrieval, and you may be losing clicks in the future too.

And none of it is a guarantee

What happens when crawling your content matters more to an AI company than being a “good citizen”?

Declaring a user agent gives people the chance to block it.

That's a fair exchange, but Grok (and more recently Muse) has been seen rotating through user agents and IP addresses until it gets in.

Mitigation is mostly about making it more trouble than it's worth - frustrating the bad actor until they give up. Not cast-iron blocking, similar to what Google is doing with SERP scraping

Dixon Jones said it best.

The user agent most responsible for unwanted crawling is often just Chrome.

A user that looks like you or I.

04

Then agents arrive

AI working on our behalf

Crawlers declare themselves (sometimes). Agents may not.

A crawler generally:

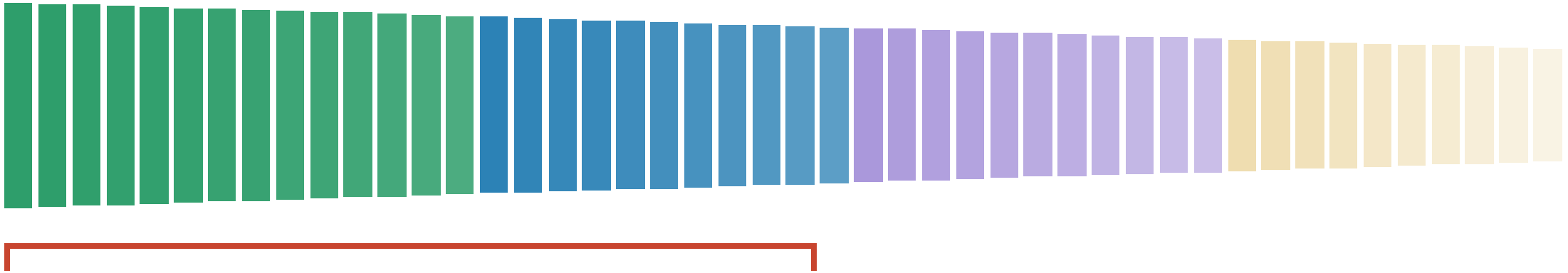
- Identifies itself
- Behaves predictably
- Can be reasoned about

An agent might:

- Be completely anonymous
- Crawl on behalf of a user
- Fire up a real browser and click through like one
- Actively try to work around bot mitigation

Headful Chromium, can look like a person even when it isn't one.

Identifiable, then not



Declared crawler

Identifies itself & behaves predictably.

Agent

May declare itself, or may not. Sometimes browser-like too

A user - or something acting as one

Indistinguishable unless you are deploying more tools to monitor behaviour

Your analytics can’t reliably tell the difference (yet).

Agentic browsing can fire web analytics. It puts data into your platforms that makes bot sessions look like real users - except they don't navigate like real users.

Business-critical information starts being contaminated with agent traffic. That is not ideal.

The cost of blocking the wrong bot

Mitigation can backfire too

CAPTCHAs and proof of work add friction for bots - and for legitimate humans, and for legitimate agents working for them.

Blocking a bot isn't always intentional

Someone asks for a price comparison on trainers.

Site 1	Site 2	Site 3	YOUR SITE	Site 5	Site 6
allowed	allowed	allowed	blocked	allowed	allowed

The answer the user sees

Site 1	Site 2	Site 3	Site 5	Site 6	—

You are not *just* beaten, completely absent.

05

From block list to bot policy

What 2026 requires

Thinking about bot policies rather than just block lists

01 Who is accessing our content?

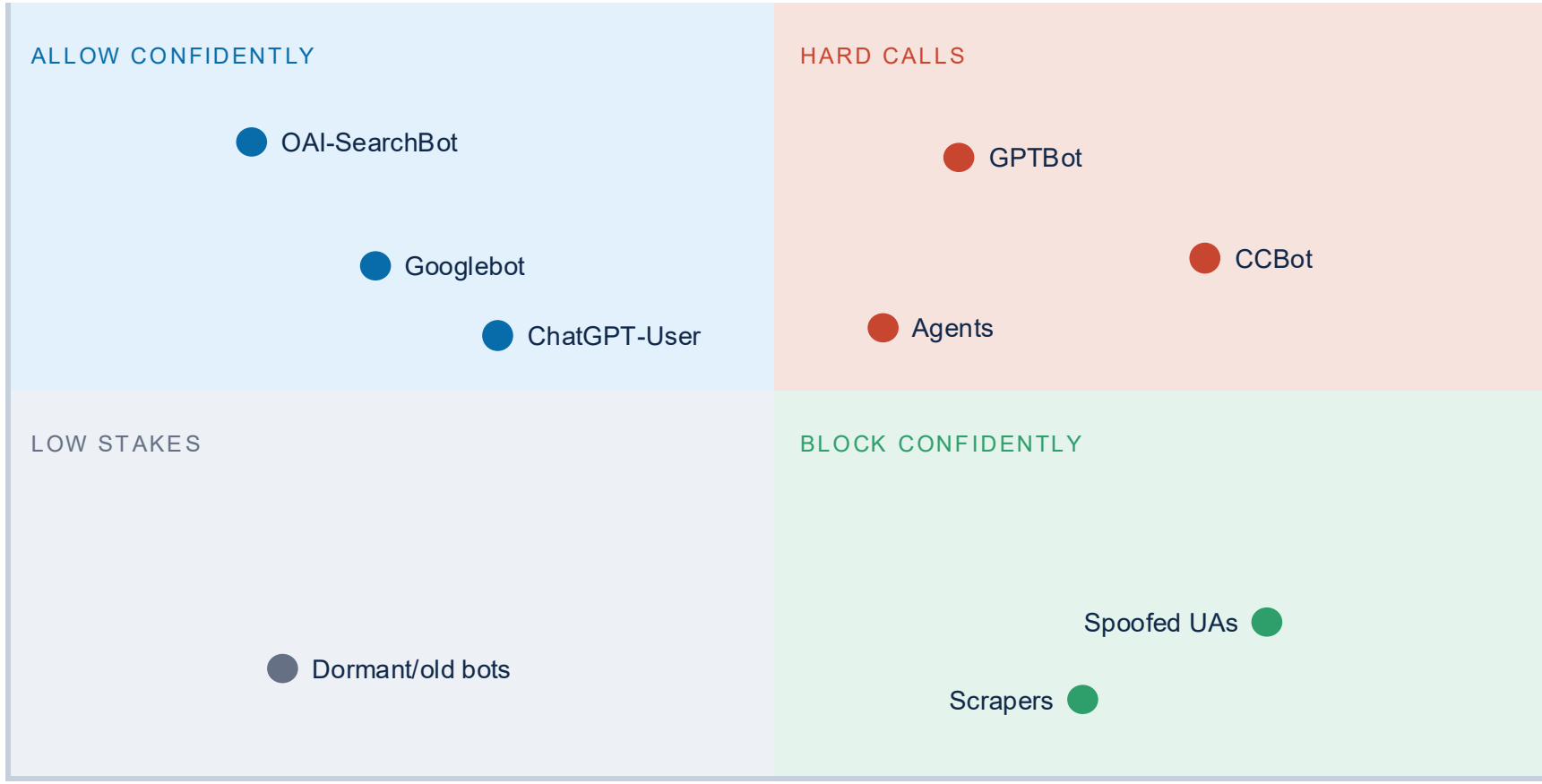
02 Why are they doing it?

03 What are the consequences for not allowing them?

04 For every block or allow, what outcome are we choosing?

Some of these outcomes will be bad for somebody. That's EXACTLY why we're documenting this and consulting the business, the stakes are so much higher.

Making tough decisions



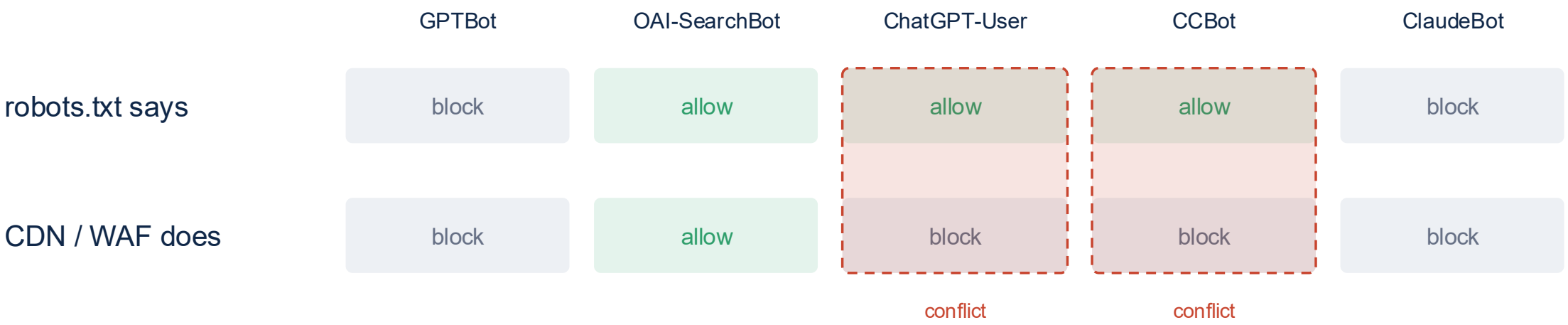
Move the dots to match your own business!

“Our brand won't appear in model memory because we blocked a bot nobody had heard of.”

That is a conversation people want to have.

They may well come around to your way of thinking, if you can make it explicit for them.

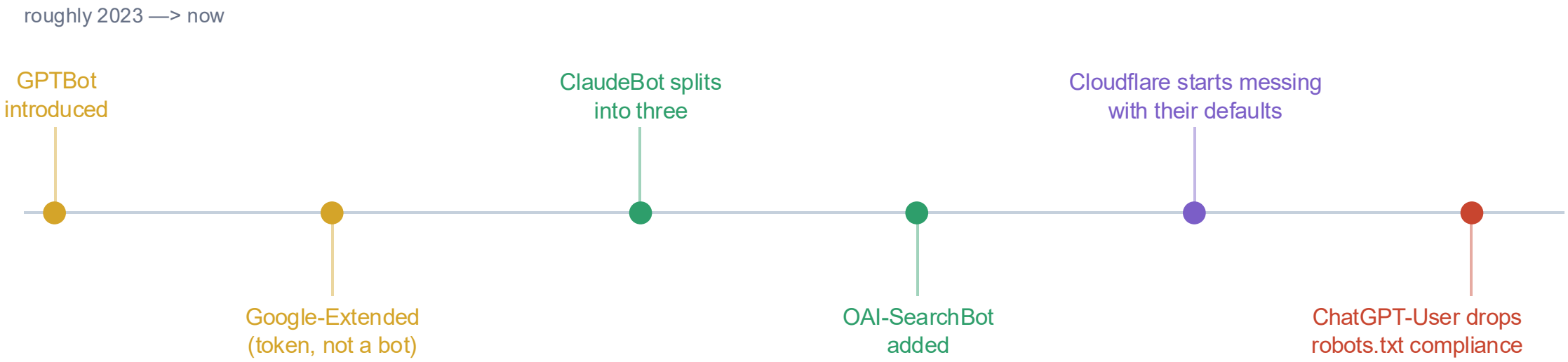
Robots.txt and CDNs can disagree too



Where the two disagree, the CDN wins and your robots.txt is undermined

What’s worse… the person who wrote one rule usually isn't the person who changed the other.

We have A LOT more to keep up with now too

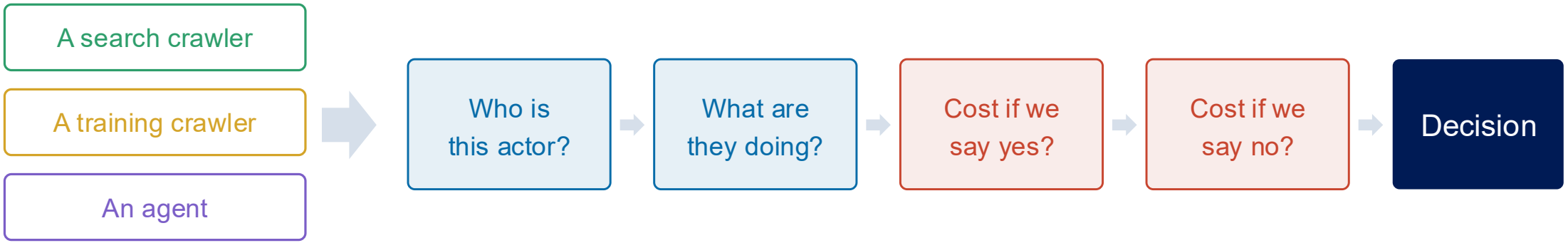


Not one of these was announced to the SEO team –
unless you were specifically looking

Then keep it aligned *and* keep it moving

Align the rules	Your robots.txt and whatever runs on the CDN or WAF have to reflect the same policy. Don't undermine one with the other.
Treat it as constant a work-in-progress	It isn't monolithic, providers, agents and user agents change constantly.
Assume you won't be told	You will not always be kept in the loop when something changes — so put the tools and processes in to catch it.

The policy process we need to be running through



Bot management is no longer “is this bot good or bad?”

The questions are now:

- Who is this actor?
- What are they doing?
- What are the costs of them participating?
- And what are the costs if they don't?

I'd now consider virtually everything here part of my remit.

Shaping how bots and search providers access and use my content.

I suggest 2026 is the year you do this too.

Thank you

@chrisgreenseo - torquepartnership.com