



Bot Management

Learnings from Bauer Media Group

Stefan Betzold
CPMO Bauer Media Group - Digital Publishing

Automated Traffic Surpassing Human Traffic

Data from Imperva Bot Report

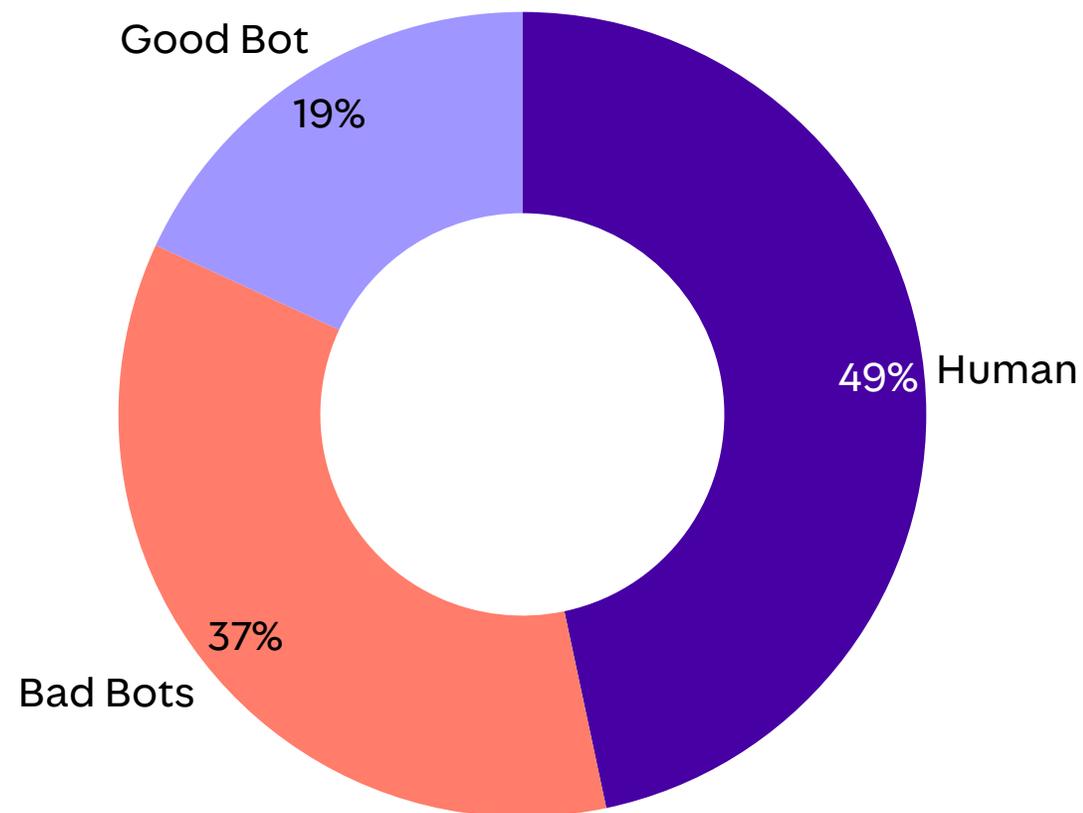
For the first time in a decade, automated traffic surpassed human activity.

Bot traffic is not new to us: Search-, SEO- and other crawlers exist for decades.

Growth is largely driven by rapid adoption of AI and LLMs.

Bad bot activity has risen for the 6th consecutive year, with malicious bots now making up 37%.

For us in **Publishing**, we have never really cared about bot protection in the past. We have to change this now...



Global Internet Traffic
Imperva Bot Report 2025

imperva



Bot traffic surpassing human traffic

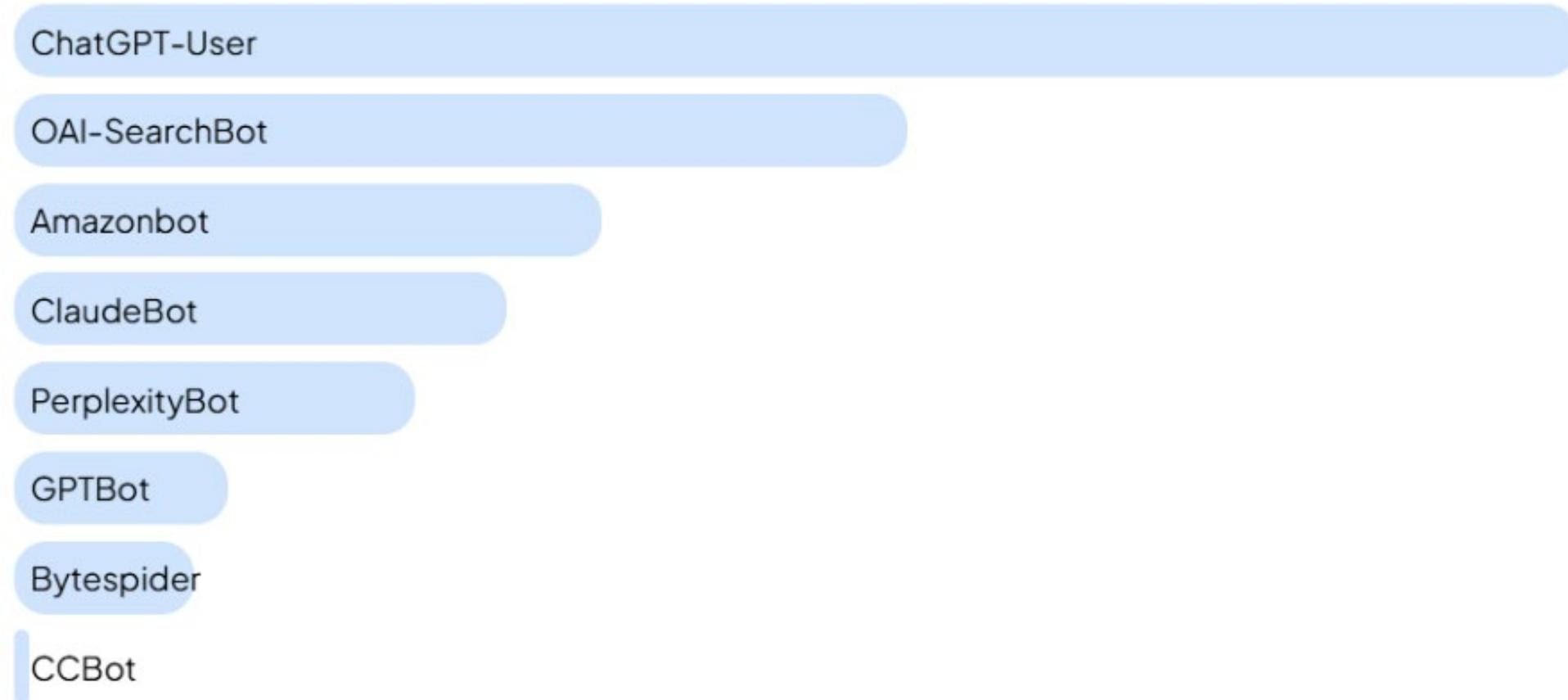
example from one Bauer Media brand in Q4



ChatGPT dominating with crawls from AI bot access

example from one Bauer Media brand in Q4

Breakdown by bot



Biggest growth coming from smaller companies

example from one Bauer Media brand in Q4

1.3M

AI bot scrapes

+32.1% ↗

from previous

AI bots made **1.3M requests** to your website, **up 32.1%** from the previous



ClaudeBot

Anthropic

+61.8% ↗

from previous

#4



PerplexityBot

Perplexity

+230.5% ↗

from previous

#5



Clickthrough Rates from AI platforms on micro-level

example from one Bauer Media brand in Q4

Clickthrough Rates

How many human visitors AI platforms are sending you.

AI Platform	Referers	Scrapes	<u>Clickthrough Rate</u>
 OpenAI (ChatGPT)	253	1.4M	0%
 Perplexity	107	153.2K	0.1%

????!!!

Not a fair deal: 500:1 to 2.800:1 crawls per referral

example from one Bauer Media brand in Q4

Referral traffic

AI companies brought **134 referrals** over the last month, **up 100.0%** from the previous period. **OpenAI** was the top referrer with **63 referrals**.

Scrape-to-referral ratio

How often AI bot scrapes result in a human visit.

134

Total AI referrals

361.4K

Total scrapes

2696:1

For every 2696 scrapes you get 1 AI referral



Top referrers

AI companies sending the most traffic to your site.

OpenAI

63 referrals

Perplexity

42

Microsoft

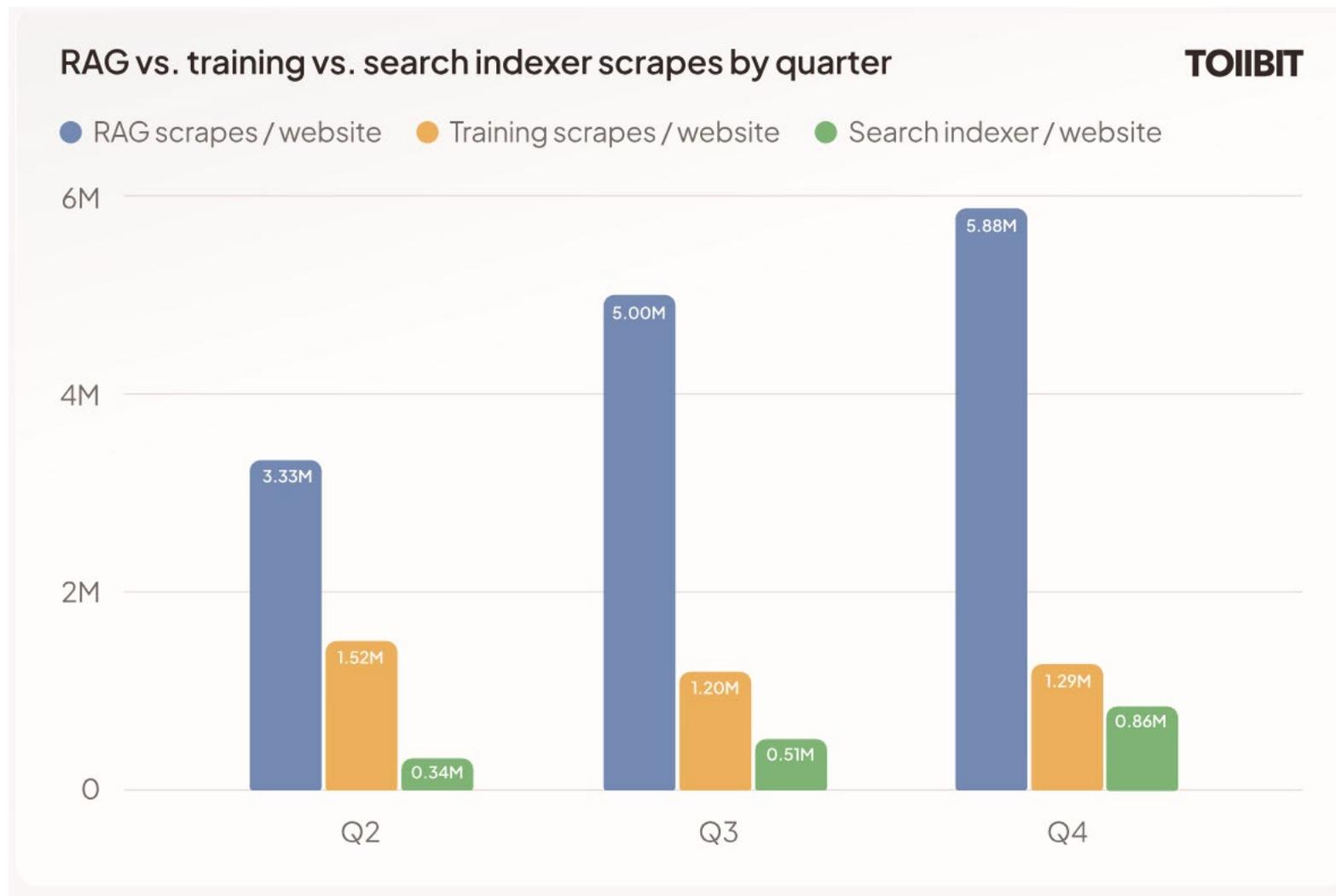
28

Anthropic

1

RAG crawling massively increasing against Training

data from Tollbit state of the bot report



In Q4, the average scrapes per page for a RAG bot was ~10× more than for a Training bot



Robots.txt – is only a guidance and often ignored

example from one Bauer Media brand in Q4

Bots bypassing

AI bots ignored your robots.txt **51.2K** times over the last month. **Meta** bypassed the most with **43.0K** attempts.

[See all bypassing data](#) →

AI bots ignoring your robots.txt

AI bots blocked in your robots.txt are still accessing your content. [Set up bot payroll now](#)

51.2K

times listed AI bots accessed your content last week

Top bypassing companies

The AI companies that ignored your robots.txt the most

Meta	43.0K
OpenAI	7.7K
ByteDance	509
Perplexity	21
Timpi	12



Publisher adoption of robots.txt

Miso.AI global robots analysis

22

average robots.txt
publisher block list

Only 40% had bot block,
42% didn't block any bots.

43

Bauer Media
Robots.txt
blocklist

2.200

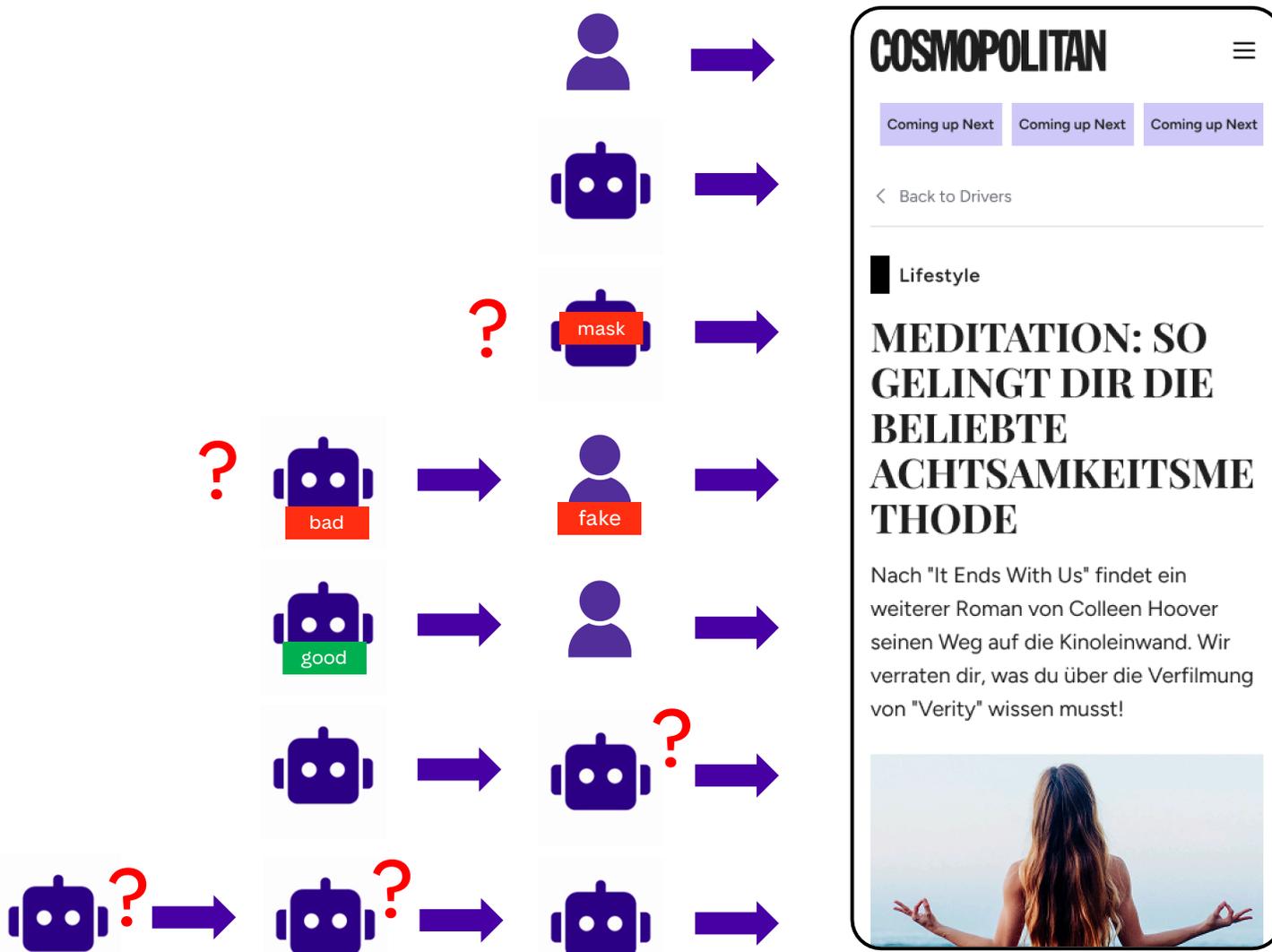
Bots accessing
publisher websites

Source: Miso's Global Robots.txt Analysis (Oct 2025)
11.500 publishers analyzed.

miso



And it's not about simple crawling bots any more!



The growing complexity of non-human access:

Direct crawling and scraping with identification (e.g. user agent)

Masqued crawling undeclared or using fake user agents

Third-party scraping services for outsourcing the scraping process

Residential IP proxies, presenting as ordinary user traffic, that match human audience.

Circumvention services that evade bot detection / IP protection measures.

Cloud-based headless browsers from new AI browsers, often used with a residential IP proxy, making the request appear to come from a normal human user.



Easier than ever to order an AI crawling agent to extract contents and data from any website

apify Product Solutions Developers Resources Pricing Contact sales [Go to Console](#)

Get real-time web data for your AI

Apify Actors scrape up-to-date web data from any website for AI apps and agents, social media monitoring, competitive intelligence, lead generation, and market research.

Extract products from Amazon

TikTok Scraper
clockworks/tiktok-scraper

Extract data from TikTok videos, hashtags, and users. Use URLs or search queries to scrape...

Clockworks 125K 4.7 (197)

Google Maps Scraper
compass/crawler-google

Extract data from thousands of locations and businesses, including...

Compass

Website Content Crawler
apify/website-content-crawler

Crawl websites and extract text content to feed AI models, LLM applications, vector databases, or...

Apify 100K 4.5 (165)

Amazon Scraper
junglelee/free-amazon-pr

Gets you product data from Amazon API. Scrapes and downloads...

Junglelee

Browse 10,000+

apify Product Solutions Developers Resources Pricing Contact sales [Go to Console](#)

Search: paywall

All categories All pricing models All developers Most relevant 76 Actors

Bloomberg Article Scraper
jamie_tran/bloomberg-article-scraper

The Bloomberg Article Scraper allows you to extract the full internal data of Bloomberg Articles, bypassing the paywalls, and returning a full complete JSON of the...

Jamie 4.0 (1) 8

Smart Article Extractor
lukaskrivka/article-extractor-smart

Smart Article Extractor extracts articles from any scientific, academic, or news website with just one click. The extractor crawls the whole website and...

Lukáš Křivka 4.9 (7) 6.8K

Bloomberg News Scraper
romy/bloomberg-news-scraper

Bloomberg News Scraper is an advanced scraper that allows you to access and extract content from Bloomberg News, even for articles that usually require ...

Romy 5.0 (1) 274

Fast News Scraper
timgreen/fast-news-scraper

Extract full article text and metadata from popular news sites like The New York Times, AP News, Reuters, CNBC, NPR, and Wired. Scrape thousands of articles l...

Tim Green 1.1 (3) 523

Ultimate Screenshot
dz_omar/ultimate-screenshot

Capture ANY website as HD screenshots, videos, or PDFs! 100+ device presets (iPhone, Android, tablets). Perfect for web scraping, monitoring, testing ...

FlowExtract API 5.0 (8) 225

Substack Leaderboard Scraper
easyapi/substack-leaderboard-scraper

Scrape detailed publication data from Substack leaderboards. Get comprehensive insights about top newsletters including subscriber counts, pricing, autho...

EasyApi 5.0 (1) 30

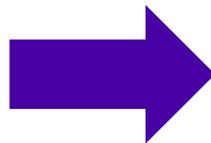
Tumblr Scraper
easyapi/substack-publications-scraper

Substack Publications Scraper
easyapi/substack-publications-scraper

Screenshot
dz_omar/screenshot



Human access: We implemented granular access and data control for Human traffic with TCF framework



Auto
ZEITUNG

Zur Verbesserung und Finanzierung unseres Webangebots erheben und verarbeiten wir und unsere 131 **Partner** personenbezogene Daten, indem wir bspw. persönliche Identifikationsmerkmale wie IP-Adressen und Gerätekennungen nutzen oder Cookies auf Ihrem Endgerät speichern.

Die Datenverarbeitung erfolgt dabei zu den folgenden Zwecken: Speichern von oder Zugriff auf Informationen auf einem Endgerät; Personalisierte Werbung und Inhalte, Messung von Werbeleistung und der Performance von Inhalten, Zielgruppenforschung sowie Entwicklung und

Einwilligen und weiter

Für 2,99 €/Monat abonnieren

Datenschutzmanager aufrufen

[Mit PUR-Abo anmelden](#) [Impressum](#) [Datenschutz](#)
[Datenschutz \(PUR\)](#) [AGB](#)

anderer nicht einwilligungspflichtiger Gründe einsetzen.

Zustimmen

Drittanbieter-Inhalte in redaktionellen Inhalten ▼

Ablehnen

Zustimmen

Funktionale Verwendungszwecke ▼

Zusatzfunktionen ▼

We also work with some partners on the basis of legitimate interest and other legal bases without consent.

[Manage Settings](#)

Hier finden Sie eine Übersicht aller Technologieanbieter,

Allen zustimmen

Ausgewählten zustimmen

Zurück

[Impressum](#) [Datenschutz](#)

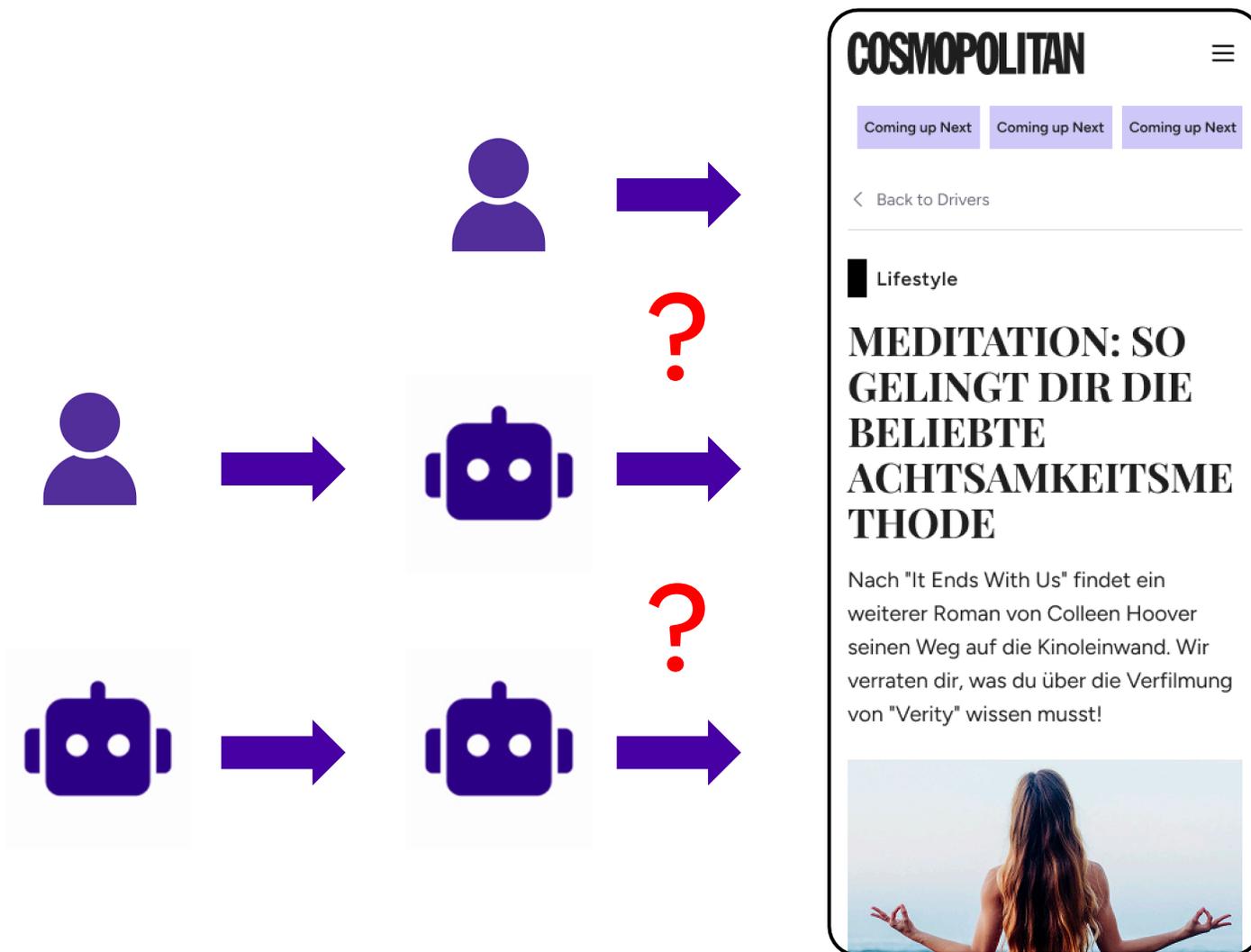
Vendor list



Purpose control
(TCF framework)



Non-Human access: We cannot identify and control bot and agent traffic



No Vendor identification

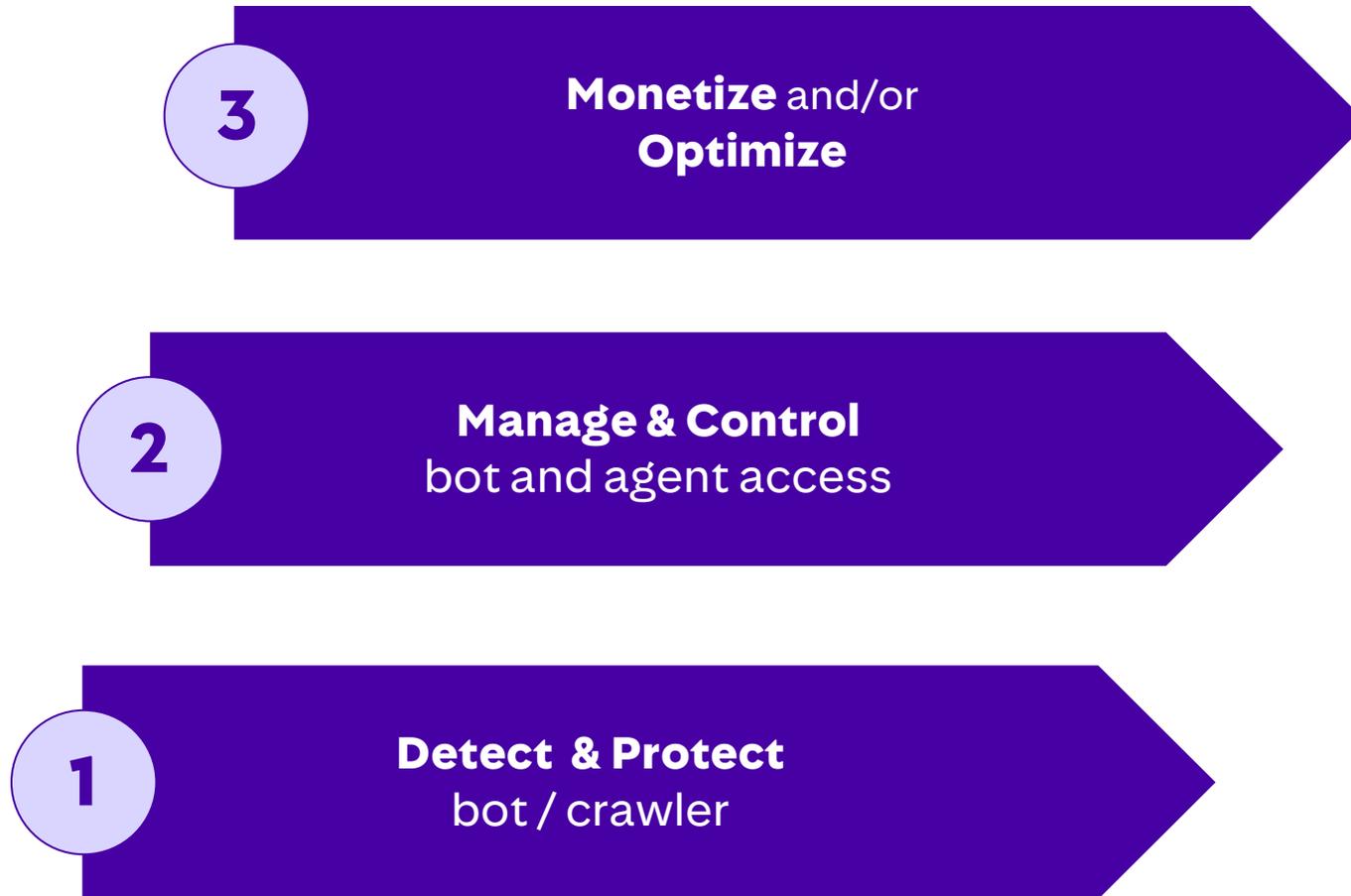
- No standards for bot / agent identification
- Bots not respecting robots.txt
- Undisclosed scraping / headless browsers
- Data brokers and third-party scraping outfits

No purpose control

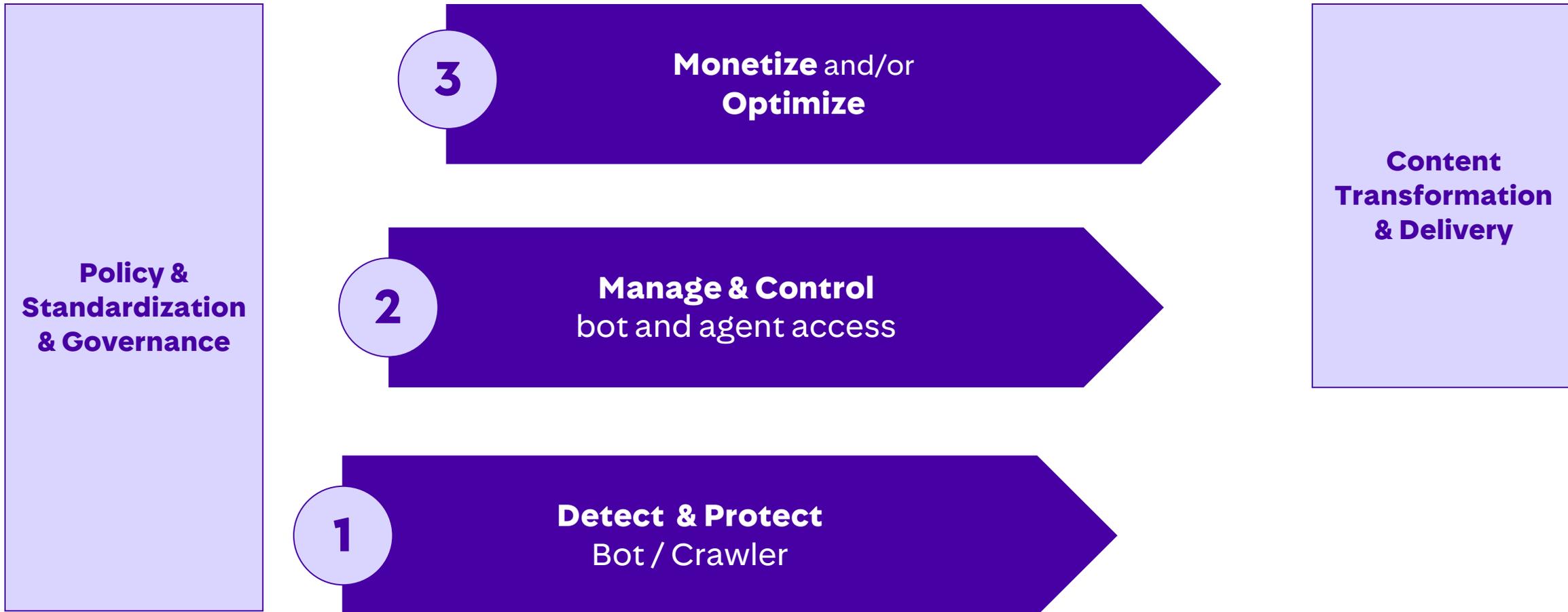
- No differentiation of e.g. Search Index, RAG, training, transaction
- Googlebot runs search and RAG for AI/O and AI mode
- Good actor agents cannot be differentiated from bad actors



What we should do



What we should do



Detect non-human access and protect from unwanted crawling

1

Detect & Protect

Detect and classify automated traffic (AI crawlers, scrapers, agents) vs. human users. Invest in bot protection (before you start thinking about monetization)

Action plan:

- Update and actively manage your robots.txt file
- Measure and analyze bot access and monitor closely (add bot-specific reporting to business dashboards)
- Evaluate bot protection tools: via CDN partner or 3rd party solution
- Start protecting! Turn on edge-level bot classification

Partnering options:

- CDN / Edge security: Cloudflare, Akamai, Fastly,...
- Specialist bot mitigation: Centinel Analytica, Datadome, Human, Imperva,...

**Policy &
Standardization &
Governance**

Support the standardization efforts:

- Lobby work as part of association work (e.g. MVFP), Brussel needs to understand!
- New protocol standards: W3C TDMRep, IETF AI PREF group



Start setting rules: move from “anonymous crawling” to “permissioned access”

2

Manage & Control

Create granular controls to manage bot and agent access for different purposes.

Action plan:

- Evaluate your “highest value contents” to be protected
- Define content access tiers: **granular controls** and rule sets by
 - content type (news, evergreen, paywalled, archives...),
 - bot purpose (training, summary, grounding, display...)
 - content use categories (free / snippet / license / premium / neverAllowed)
- Provide controlled endpoints (instead of letting bots hit the human website).
- Policy, rights, and governance (the “rules” layer)

Partnering options:

- Connect to bot management / cybersecurity providers
- Emerging landscape of new players: e.g. tollbit. But a lot of room for innovation!

Standards:

- Protocol standards: auth/token access patterns, llms.txt “AI-ready guidance”, RSL



Optimize and monetize: Convert legitimate demand into revenue

3

Optimize & monetize

Prepare content transformation into a cleaned-up, for bots optimized format. Transform bot/agent traffic from cost into revenue; negotiate, meter, and license access.

Action plan:

- Reduce cost-to-serve with content transformation („reduce crawling costs“)
 - Structured feeds/endpoints for licensed access (e.g. MCP)
 - Clean content cache” for agents (faster, cheaper, controlled)
- Decide monetization routes: Direct deals, vs. Marketplace vs. Pay-per-use models
- Package your offer in commercial products (easy to sell + easy to buy), e.g.: “Licensed RAG / grounding access” (usage-based) vs. “Summaries + attribution” bundle“ vs. Full-text display” premium tier (strict terms)
- Explore and test new user propositions in conversational web.

Partnering options:

- Marketplaces: Tollbit, CloudFlare pay-per-crawl, Microsoft PCM, Prorata
- Make your own deal with LLMs, but this will be tough...

Standards:

- IAB Tech Lab ComP, RSL (really simple licensing), NIweb

Content Transformation & Delivery



What's next?

- **For AI companies**
“Treat content as strategic infrastructure - not free raw material.”
- **For publishers & creators**
“Block and manage AI bots, demand transparency, and build licensing standards.”
- **For regulators**
“Create fair, simple rules that reward high-quality, diverse content - before it disappears.”



Thank you!

