AI: crawlers hungry for data that overload your servers and leave your sites slow or inaccessible

Bernard Sfez - 2025-11-01 19:42

In the breakneck race to hoover up data, AI models are unleashing armies of bots that crawl the web nonstop—often ignoring rules like robots.txt. The result? Siphoned bandwidth, strained servers, and pages that slow to a crawl or go offline—an abnormal, excessive load that hurts publishers, businesses, and users alike. As *The Register* notes, more and more voices are calling out bots that send no traffic back yet rack up infrastructure costs and leave teams overwhelmed.

At OpenSource Solutions, we take a clear, measurable approach. We continuously analyze traffic logs to spot overconsumption, correlate signals across multiple servers to tell trustworthy crawlers (search engines) from aggressive AI scrapers, and roll out proactive defenses—quotas, rate limiting, IP allowlists, bot traps, even GeoIP filtering—to protect performance without blocking the indexing that's essential for SEO. Let's break it down.

Why AI crawlers are now a problem

The boom in AI models has triggered a frantic grab for content. Unlike traditional search engines—which aim to index pages and send traffic back—many AI-related bots bring you no visitors: they vacuum up pages, APIs, and even media assets at scale to train models, offloading the cost (bandwidth, CPU, storage, support) onto publishers and businesses. In recent days, alerts and investigations have multiplied (The Register, Cloudflare, Ars Technica, Search Engine Land).

Beyond simply ignoring your robots.txt directives, these crawlers "don't play by the rules" in many ways:

- Opaque or fake user agents: generic strings ("Mozilla/5.0", "Python-requests…") or impersonating a legitimate UA.
- Network cloaking: IP rotation, multiple hosts, proxies, varied ASNs to evade blocking.
- Excessive pace: bursts of parallel requests, no backoff, ignoring overload signals (429/Retry-After).
- Standards workarounds: disregarding rate limits, crawl-delay, sitemaps, and exclusion zones.
- Aggressive discovery tactics: dictionary/parameter probing, directory scans, following transient links, fetching internal APIs or exports not intended for the public.
- Cache bypassing: tweaking headers and parameters to force cache misses and multiply server cost.
- "Human-like" automation: headless browsers and scripts that execute JS and simulate interactions to slip past simple barriers.

This asymmetry creates cascading effects—bandwidth and server budgets drained, pages slower or even unavailable for days, noisy analytics, and overwhelmed support teams—with no business upside. At scale, this isn't mere robotic discourtesy; it's an operational risk (SLO, costs), a marketing risk (SEO/UX degradation), and sometimes a contractual risk (terms of use/API, content rights).
The challenge isn't to ban all robots, but to separate useful crawling (the kind that helps your search visibility) from opportunistic scraping, and then regulate the latter.

Symptoms and diagnosis

What you and your visitors might notice

- The site gets slow: pages take many seconds—or even minutes—to load.
- The site goes down intermittently: error messages, blank pages, or endless spinners.
- Actions that “stick”: shopping cart won’t open, search stops responding, forms won’t submit.
- Outage spikes at the “wrong” times: late at night or early morning when you expect little traffic.
- Unexplained cost increases: hosting/CDN bill suddenly higher.
- Complaints from customers: “It’s super slow,” “I get an error,” “I can’t edit my pages.”

If you’re seeing at least two of these, your site may be getting “eaten” by overly greedy AI bots.

Why this happens

Automated robots download your pages far more often—and faster—than any human would.
Result: they hog the line (bandwidth) and overload the machine (server), slowing everyone down and sometimes cutting access entirely.

What to do right now

At minimum, run these checks:

- Try from a different network (not just another workstation at the office). From home, or via your mobile without Wi-Fi (4G/5G), to rule out a local issue.
- Post a clear message to users when you detect problems (“We’re investigating, thank you for your patience”). This reduces support load and buys engineers time.
- Watch over 24–48 hours to see if slowdowns/errors come in waves (and note the times).

At OpenSource Solutions we can run the initial checks with you, or fully on your behalf—and handle incidents while keeping legitimate search engines connected.

More technical checks

Read this if you have technical skills and access to your host/CDN, metrics (e.g., Zabbix, Matomo) and access logs.

Typical technical indicators (in hosting metrics, monitoring tools, and logs)

- Request spikes at off-hours, without a rise in real visitors.
- Bursts on deep URLs (archives, filters, heavy images).
- Generic or shifting User-Agent (“Mozilla/5.0”, Python scripts, scraping tools).
- IP rotation (cloud hosts, varied countries), sometimes several per minute.
- Cache bypassed: same pages re-requested with useless parameters (?v=…, ?rand=…).
- 429/5xx errors climbing during bursts, then dropping back down.

Quick mini checklist

- Top User-Agents and IPs/ASNs over the last 24–72 hours.
- Cache-miss rate and most expensive pages (images, JSON, search).
- Time windows of spikes vs. human traffic.
- Volume of 429/5xx errors during spikes.

If all this resonates but it’s not your day-to-day, no worries: our team at OpenSource Solution can run these checks for you (guided diagnosis or turnkey).

Tangible impacts: technical, business, legal

Simply explained overview

- Poor customer experience: sluggish pages, failed payments, forms that spin—and visitors who leave.

- Lower revenue: fewer conversions (sales, contact requests), more cart abandonment.
- Brand damage: negative reviews, perception of an “unreliable site.”
- Exploding costs: hosting/CDN and customer support bills climb with no benefit.
- Compliance risks: unwanted content extraction, violations of terms of use, accidental exposure of poorly scoped public data.

More technical view to quantify impact

- Technical / SLO: higher TTFB, degraded Core Web Vitals, spikes in 5xx/429, CPU/I/O saturation, clogged job queues.
- Business: lower conversion rate, higher bounce rate, shorter sessions during bursts; for e-commerce, direct impact on hourly revenue.
- Financial: extra CDN egress, storage (logs, caches), compute (runaway autoscaling), support time L2/L3.
- Legal / Contractual: breaches of your ToS/API terms, reuse of content without permission, GDPR obligations if endpoints expose too much public data.

Our response: implementing key components

Smart monitoring (Zabbix + Matomo)

- Pragmatic alert thresholds (e.g., TTFB > 1.2 s for 5 min, 5xx errors > 2%, request spikes without a rise in human visitors).
- Multi-channel alerts (email, SMS, etc.), prioritized by business impact (cart, checkout funnel, key pages).
- Zabbix ↔ Matomo correlation: if load rises without human traffic, we trigger anti-bot defenses.

Real-time protection (WAF + Fail2ban)

- Targeted slowing (rate limiting) as soon as a robot exceeds “normal” cadence.
- Automatic blocking of suspicious IPs/ASNs via Fail2ban (greylist first, then blacklist on repeat).
- Dynamic WAF rules: cut typical scraper behaviors (cache bypass, heavy-image floods, URL scanning).
- Discreet honeypages to trap bots that ignore rules and accelerate their exclusion.

Fine-grained handling of good crawlers (Virtualmin + WAF)

- Verified allowlist (official reverse DNS/IPs) for Google, Bing, Applebot, etc.
- Gentle quotas for legitimate engines (they pass, but don’t saturate).
- Sitemaps & priorities: steer indexing toward pages that matter to preserve/boost SEO.

Smooth operations (Virtualmin)

- Targeted restarts and load spreading (without taking the site down).
- Caching and PHP-FPM tuning adapted to your seasonality.
- Incident playbooks: who does what, in what order, when a spike hits.

Expected results, what this changes for you : Pages that stay fast even during bot spikes. Controlled bills (avoided bandwidth/egress), fewer tickets and support requests about a “slow site,” and more useful visits or conversions at key moments. SEO is preserved: good engines continue indexing what matters.
Examples of alerts & automatic actions

- Alert: “5xx error rate > 3% for 3 min on /search” → Action: throttle offending IPs + force caching of lightweight responses.
- Alert: “+300% requests for original images in 10 min” → Action: progressive delay and waterfall block on the source ASN.
- Alert: “Request spikes without a Matomo (human) increase” → Action: Fail2ban switches to aggressive mode and the WAF tightens scraping rules.

Conclusion: a useful Web is a fair Web

The picture is now public: as multiple sources have documented, waves of AI-related crawlers trigger traffic spikes, degrade performance, and stick publishers with the bill—without sending traffic or value back.

In response, major infrastructure players have begun shutting off unauthorized AI crawlers by default and launching “Pay Per Crawl” mechanisms to give publishers a say—and, where appropriate, compensation when robots exploit their content. (Cloudflare, Wired, The Verge, Nieman Lab.)

Tomorrow, these practices may well be written into law: clear bot identification, explicit consent for certain AI uses, liability for damages—and even contractual fees for large-scale machine access. The debate is moving as incidents pile up, like the Perplexity case accused of “stealth crawling” that circumvents access rules. (Business Insider)

Beyond the tech, this is about balance: let through what brings value (indexing, discovery, traffic), and regulate what takes without giving back (opportunistic scraping, abusive load). The webmaster’s craft returns to center stage as the conductor: publish useful information, define a clear indexing framework, monitor with care, and engage with legitimate robots. Our role is to help you keep that balance: protect your visitors, your costs, and your reputation—without breaking your SEO.

Let’s talk: we can run the first diagnosis with you (or for you) and implement the right protections, at your pace.