Managen und Monetarisieren von KI-Bot-Besuchen

KI-Scraper sind ein struktureller Faktor im Medienwandel. Für Publisher, Vermarkter und Plattformbetreiber ist Transparenz über Art, Umfang und Entwicklung der Bot- und Scraper-Aktivitäten die Grundvoraussetzung für fundierte Entscheidungen über den Umgang mit ihren Inhalten, deren Schutz und Monetarisierung. Sie greifen in großem Umfang auf Inhalte zu, ohne selbst Traffic oder Erlöse zurückzuführen, und entkoppeln damit Reichweite zunehmend von Monetarisierung. Gleichzeitig steigen Frequenz und Volumen dieser Zugriffe deutlich, was sowohl Infrastrukturkosten erhöht als auch den Kontrollverlust über eigene Inhalte verstärkt.Doch wie können aus transparenten Informationen konkrete Steuerungsoptionen abgeleitet werden? Welche Rolle spielt die Robots.txt heute noch, wo liegt ihre Grenze gegenüber KI-Scrapern? Welche technischen Optionen zur automatischen Blockierung und welche Risiken müssen dabei beachtet werden?

Diese Fragen werden nachfolgend beantwortet. Ziel dieses Papiers ist es zu zeigen, wie Publisher KI-Bot-Aktivitäten auf ihren Webseiten besser managen und monetarisieren können – und dabei gleichzeitig Sichtbarkeit, Vermarktung und Compliance gewährleisten.

Passive Blockierung über die Robots.txt: Bot-Transparenz im Quick-Check

Die robots.txt basiert auf dem sogenannten Robots Exclusion Protocol (REP), das bereits Mitte der 1990er Jahre als informeller Standard zur Steuerung von Webcrawlern entstand und später weiter formalisiert wurde. Es definiert grundlegende Regeln wie Allow und Disallow, anhand derer Website-Betreiber festlegen können, welche Bereiche einer Seite von Crawlern abgerufen werden dürfen.

Eine entsprechende Domain kann etwa so aussehen:
beispielwebseite.de/robots.txt

Die zugrundeliegende Syntax ist wie folgt strukturiert:
User-agent: *
Disallow: /private/
Allow: /public/

Erst durch das aktive Mitwirken von Google und anderen großen Akteuren sowie die spätere Standardisierung im W3C fand die robots.txt breite Akzeptanz unter Publishern. Heute halten sich die meisten klassischen Suchmaschinen-Crawler an diesen Standard und respektieren die in der robots.txt definierten Regeln.

Crawling durch KI-Scraper

Der Aufstieg von generativen KI-Modellen wie ChatGPT, Gemini, Claude oder LLaMA hat eine neue Klasse von Web-Crawlern hervorgebracht: **KI-Scraper**. Sie durchsuchen die Websites nach Inhalten und Daten und nutzen diese für verschiedenen Zwecke. Die nachfolgende Tabelle schlüsselt die verschiedenen Nutzungsarten von KI-Scrapern auf und zeigt, wie sich das Bot-Verhalten erkennen lässt.

Übersicht: Gängige Nutzungsarten der Publisher Inhalte

Nutzungs-Kategorie	Kurzbeschreibung	Typisches Bot-Verhalten
Training	Modell lernt statistische Muster aus großen Content-Mengen (Pre-Training/ Finetuning).	Hohe Volumina, breite Coverage, selten wiederkehrend.
RAG	Retrieval-Augmented Generation: Snippets werden zur Anfragezeit in den Kontext gepackt, um Antworten zu verbessern. Siehe auch RAG Whitepaper[1] vom BVDW.	Regelmäßige, gezielte Fetches; oft snippet-basiert; manchmal über Drittdienste/Indizes.
Grounding	Antwort wird an vertrauenswürdige Quellen "geankert", um Halluzinationen zu reduzieren.	Abruf und explizite Referenz/ Belegführung; teils mit strenger Quellenpolitik.
Agent-View	Agent beobachtet/plant: sammelt Optionen, Preise, Verfügbarkeiten für spätere Aktionen.	Session-ähnliche Zugriffe, Navigation über mehrere Seiten, Vergleichsmuster.
Agent-Actions	Agent führt Handlung aus: Reservierung, Bestellung, Formular, Kündigung etc. (browserbasiert möglich). Siehe auch KI-Agent Definition[2] vom BVDW.	Login/Checkout-Flows, Form-Posts, Bestätigungsseiten; braucht Fehlertoleranz.
Indexing & Embeddings	Content wird extrahiert, segmentiert, vektorisiert (Embeddings) und in Such-/ RAG-Indizes abgelegt.	Viele "GET"-Requests, Fokus auf Text, Überschriften, Listen; teils Headless Browser.
Display	Systeme nutzen die gecrawlten Inhalte um die „besten Passagen" in die Antorten einzubinden. Zum Beispiel für Kurzzusammenfassungen, Lokalisieren oder Zitate etc.	Abruf ganzer Seiten, Abruf und Prüfung von Metadaten, URLs etc
Entity & Knowledge Graphs	Extraktion von Entitäten/Beziehungen (Person/Organisation/Produkt/Ort) für Graphen.	Fokus auf strukturierte Daten, Tabellen, wiederkehrende Muster.
Evaluation / Benchmarking	Content dient als Testkorpus: Faktenchecks, QA-Benchmarks, Retrieval-Qualität.	Periodische Crawls derselben Seiten, um Drift zu messen.

Publisher stehen nun vor der Frage, wie sie diese Zugriffe kontrollieren, gezielt einschränken oder vollständig unterbinden können, ohne den eigenen Traffic oder SEO-Wert zu gefährden.

Der Trend zeigt, dass immer mehr Publisher KI-Crawler explizit in der Robots.txt listen. Allerdings halten sich diese nicht immer an den Standard und scrapen die Inhalte der Webseite trotz vorhandenem Eintrag in der Robots.txt.

In der Robots.txt selbst kann nicht unterschieden werden, für welchen Zweck das Crawling oder Scraping erfolgt. Man muss also genau darauf achten, welche Crawler zugelassen oder geblockt werden, um unerwünschte Nebeneffekte, wie z.B. Traffic-Verluste oder Search Rank-Verluste zu vermeiden.
https://radar.cloudflare.com/ai-insights öffentlich einsehbar.

1 https://www.bvdw.org/news-und-publikationen/retrieval-augmented-generation-rag/
2 https://www.bvdw.org/news-und-publikationen/ki-agenten-definition/

Übersicht der bekanntesten KI-Crawler[3]

Anbieter	Crawler / User-Agent	Zweck	robots.txt Block
OpenAI	GPTBot	LLM-Training	GPTBot
OpenAI	ChatGPT-User	Real-time web fetching	ChatGPT-User
OpenAI	CCBot	Web archive (für Training)	CCBot
Anthropic	Claude-Web-Agent	Websuch	Claude-Web
Perplexity	PerplexityBot	Suche/Index	PerplexityBot
Google	Googlebot	Suchindex	Googlebot
Google	Gemini KI Agent	Generatve KI Antworten	Google-Extended
Microsoft	bingbot	Suchindex	bingbot
Microsoft	BingPreview / BingChat	Chat-based Web Fetching	BingPreview
Microsoft	Microsoft-Extended	Training Opt-Out	Microsoft-Extended
Meta	Meta-ExternalAgent	AI-Training	Meta-ExternalAgent
Meta	Facebookexternalhit	Preview Scraping	Facebookexternalhit

Eine ausführlichere Übersicht der KI-Crawler kann zudem im Cloudflare Bot Directory[4] abgerufen werden.

Die Anzahl neuer Bots steigt von Woche zu Woche. Folglich braucht es Tools (z. B. CDNs oder Bot-Management Anbieter) und Prozesse, die es Publishern ermöglichen, eine saubere und zuverlässige Kontrolle seiner Website-Zugriffe zu ermöglichen.

Aktive Blockierung von KI-Scrapern über Server- oder CDN-Regeln

Eine weitere Möglichkeit ist, bekannte Scraper via Server- oder CDN-Regeln zu blockieren bzw. deren Zugriff auf bestimmte Bereiche der Webseite einzuschränken. Viele CDNs bieten dafür Tools. Zum Beispiel Cloudflare WAF, der Akamai Bot Manager oder NGINX/Apache Regeln. Es gibt aber auch Cyber-Security- und Bot-Management-Anbieter, die sich auf die Erkennung und Blockierung von ungewünschten Bots spezialisiert haben.

Diese Tools oder Regeln identifizieren Scraper meist über deren IP-Adresse oder User Agent. Allerdings haben etliche KI-Scraper begonnen ihre Identität über die IP-Adresse oder den User Agent aktiv zu verschleiern, um nicht erkannt zu werden.

Einige Bot-Manager sind zudem dazu übergegangen selbst KI-Modelle und Machine Learning zu nutzen, um KI-Scraper zu erkennen und zu blockieren. Sie nutzen dafür Rate Limiting oder Anomaly Detection, also ungewöhnliches Verhalten von Website Traffic: Dazu zählen beispielsweise extrem hohe Seitenabrufe, fehlende JavaScript-Unterstützung, fehlende Scrolling-Möglichkeiten oder extrem kurze Seitenaufenthalte.

Dies kann sehr wirksam gegen nicht deklarierte Scraper sein und gleichzeitig die erwünschten SEO-Bots schonen. Allerdings ist der Einsatz solcher Lösungen kostenintensiv und oft nicht fehlerfrei.

3 https://radar.cloudflare.com/ai-insights öffentlich einsehbar
4 https://radar.cloudflare.com/bots/directory?category=AI_CRAWLER&kind=all

Beispiele von Bot Managern mit Rate-Limiting / Anomaly Detection:

Anbieter	Rate Limiting	Anomaly Detection
Cloudflare	x	x
Akamai	x	x
Imperva	x	x
Radware	x	x
HUMAN Security	x	x
F5/Shape	x	x
Kasada	x	x
DataDome	x	x

Welche Risiken können durch Blocking entstehen?

Je schärfer Publisher KI-Scraper blockieren, desto größer das Risiko, auch erwünschte Bots zu treffen. Wird etwa Googlebot durch Web Application Firewall-Regeln (WAF), Rate Limiting oder Bot-Scores ausgebremst, kann das Crawling leiden. Neue Inhalte werden dadurch später oder gar nicht indexiert, Rankings und die Sichtbarkeit in den Suchergebnissen können einbrechen.

Doch auch für die Monetarisierung durch Werbung bestehen Risiken. So können Ad-Verification-Anbieter, Messdienste oder Partner-Crawler ungewollt mitgeblockt werden. Das kann Kampagnenreports, Reichweitenmessung und technische Integrationen beeinträchtigen.

Auch Machine Learning-basierte Anomaly Detection kann legitimen Traffic fälschlich als verdächtig einstufen (z. B. bei viralen Artikeln oder sehr aktiven Nutzer*innen). Deshalb brauchen Publisher für jedes Bot-Blocking-Konzept saubere Whitelists, klare Ausnahmen für Suchmaschinen- und Partner-Bots sowie ein kontinuierliches Monitoring der Effekte.

Sollte eine Einschränkung von KI-Scrapern über die robots.txt, Server- sowie CDN-Regeln oder andere Blockierung nicht möglich sein, lassen sich Inhalte hinter Paywalls (Bezahlschranken) verbergen. Dies ist für viele Publisher, abhängig von ihren Geschäftsmodellen, jedoch häufig keine realistische, wirtschaftlich sinnvolle Option.

Denn: Selbst wenn Publisher das Webseiten-Scraping erfolgreich verhindern, können LLMs immer noch Daten aus alternativen Quellen beziehen und ihre Modelle mit diesen Inhalten trainieren.

Ausblick und mögliche Maßnahmen

Nachdem wir gesehen haben, wie schwer es ist, KI-Scraper allein mit Robots.txt, Server-Regeln und Bot-Managern im Griff zu behalten und welche Risiken ein zu restriktives Blocking für Googlebot, Ad-Verification Anbieter oder andere „gute Bots“ mit sich bringt, stellen sich die Fragen: Wie geht es weiter? Welche strukturellen Lösungen zeichnen sich ab, damit Publisher nicht jeden einzelnen Bot manuell jagen müssen, sondern langfristig mehr Kontrolle und auch Einnahmemöglichkeiten bekommen?

IAB Tech Lab (CoMP)

Das IAB Tech Lab entwickelt mit CoMP (Content Monetization Protocols) einen einheitlichen Standard. Dieser soll es Publishern ermöglichen, ihre Content-Assets maschinenlesbar zu beschreiben. Auf dieser Basis können KI-Anbieter die Inhalte beispielsweise über Content-Marktplätze lizenzieren. Gleichzeitig können Publisher festlegen, in welchem Umfang und zu welchen Bedingungen KI-Systeme ihre Inhalte nutzen dürfen.

Dahinter stecken zwei Kernelemente:

- **Zugriffskontrolle („Digital Paywall“)**

  Bots sollen Inhalte nicht mehr einfach „kostenlos“ abholen können, nur weil eine Seite öffentlich im Web steht. Stattdessen definiert der Publisher in einem standardisierten Format:
  - Wer darf crawlen (welcher Anbieter, welcher Bot)?
  - Wozu dürfen die Daten genutzt werden (z. B. nur für das Beantworten von Fragen, nicht für den Weiterverkauf)?
  - Unter welchen Bedingungen (z.B. nur Auszüge, keine vollständigen Artikel, ggf. zeitlich verzögert)?

- **Pay-per-Crawl / Pay-per-Use für Content-Lizenzierung**

  Auf dieser Basis lassen sich unterschiedliche **Vergütungs- und Lizensierungsmodelle entwickeln.** Das sind zum Beispiel Bot-Paywalls, bei denen jeder Crawl, jede angefragte URL oder jede Nutzungseinheit (z.B. „Antwort auf eine Nutzerfrage mit einem bestimmten Artikel als Quelle“) technisch gezählt und abgerechnet werden. Aber auch zeitbasierte Lizenzmodelle für dauerhaften Zugang zum Inhalt der Webseite.

Für Publisher heißt das: Statt reinem „Opt-out oder kostenloser Nutzung“ entsteht die Möglichkeit, Inhalte kontrolliert und lizenzbasiert für KI-Modelle zu öffnen.

Wichtig ist: CoMP ist noch in der Entwicklung. Eine erste Version des Protokolls ist für Ende März 2026 geplant mit Stand 11.3. in Public Comment. Diese wird sukzessive weiter ausgebaut. Die Richtung ist klar: weg von individuellen Sonderlösungen, hin zu einem interoperablen Standard, den viele Publisher und KI-Anbieter verstehen und automatisiert umsetzen können. Interessierte können der CoMP Arbeitsgruppe[5] im IAB Tech Lab beitreten und an der Gestaltung dieses Standards partizipieren.

CDNs und Content Marktplätze als „Durchsetzungs-Infrastruktur“

Damit solche Regeln im Alltag funktionieren, braucht es Stellen, die sie technisch durchsetzen.

Ein **CDN-Anbieter** wie Cloudflare oder Akamai sitzt wie ein „Schutzschild“ vor der Website. Jeder Aufruf, egal ob von einem Menschen oder einem Bot, läuft zuerst durch dieses Netz aus Servern und wird dort geprüft. Genau an dieser Stelle können dann CoMP-Signale und Publisher-Regeln umgesetzt werden:

- **Der CDN-Anbieter oder ein integrierter Bot-Management-Anbieter erkennt den Bot anhand der technischen Signale:**
  „Das ist Bot X von Anbieter Y, er meldet sich mit Zweck Z.“

- **Der Anbieter entscheidet:**
  - Zugriff erlauben, nur auf bestimmte Inhalte begrenzen oder komplett blockieren.
  - Zugriffe mitzählen und später abrechnen (Pay-per-Crawl / Pay-per-Use).

5 https://iabtechlab.com/working-groups/content-monetization-protocols-comp-for-ai-working-group/

Für Publisher bedeutet das, dass sie nicht selbst hunderte IP-Adressen, User-Agent-Strings und Verträge pflegen, sondern kann sich auf Policies und Reports verlassen, die zentral am Edge umgesetzt werden.

Darüber hinaus entstehen **Marktplatz-Modelle, die auch in Kombination mit CDNs oder Bot Management Anbietern eingesetzt werden können:**

- CDN oder Bot Management Anbieter erkennen einen ungewünschten Zugriff und leiten den Bot an eine "Bot Paywall" eines Marktplatz-Anbieters weiter.
- Mehrere Publisher bündeln ihre Inhalte und Regeln an einem Ort.
- KI-Anbieter greifen über einen einheitlichen technischen und vertraglichen Rahmen auf diese Inhalte zu.
- Die Plattform übernimmt Matching, Berechtigungsprüfung, technische Anbindung und Abrechnung.

Für Publisher reduziert das Komplexität. Statt vielen Einzelverträgen und technischen Integrationen gibt es einen oder wenige zentrale Partner, über die sowohl Kontrolle als auch Monetarisierung laufen. Hinzu kommt, dass kleinere Publisher auch keine Einzelverträge mit den LLMs bekommen.

Übersicht von CDN und Marktplatz-Anbietern mit Lösungen für Content Monetarisierungsmodellen:

Anbieter	Typ	Monetarisierungsansatz	Besonderheiten
Cloudflare	CDN	Pay-per-Crawl / Request-Gebühr	Globale Reichweite, integrierte AI Crawl Control
TollBit	Content Licensing Platform	Pay-per-Query / Revenue Share	Spezialisierte KI-Bot-Paywall für Publisher
Microsoft	AI Content Marketplace	Pay-per-Usage für Copilot & weitere AI-Produkte	Direkter Zugang zur Microsoft-AI-Nachfrage
ProRata	Licensing & Search Plattform	Revenue Share, attribution-basierte Auszahlung	Starke Ausrichtung auf News- und Medienpublisher
Content Bridge	Content Licensing Platform und Bot Management	Usage-basiert, Revenue Share	EU-Publisher-Fokus, Bot-Management und LLM-Lizenzen
Dappier	AI Data Marketplace	Usage-basierte Fees für Daten-Feeds	Echtzeit-APIs für LLM- / RAG-Szenarien
RSL Collective	Offener Standard	Verschiedene Modelle (z. B. Pay-per-Crawl)	Offenes Protokoll, kollektive Verhandlungsposition

Interne Governance: „AI-Bot-Policy“

Zudem braucht es intern klare Zuständigkeiten:

- Wer entscheidet, welcher KI-Bot geblockt oder zugelassen wird?
- Wie werden SEO, Redaktion, Vermarktung, IT und Legal eingebunden?
- Wie oft werden Whitelists/Blocklists und die Robots.txt überprüft?

Eine einfache, aber verbindliche **AI-Bot-Policy** hilft, Ad-hoc-Entscheidungen zu vermeiden: Sie definiert Ziele (z. B. Schutz von Premium-Inhalten, Wahrung der SEO-Sichtbarkeit), technische Leitplanken und Prozesse (z. B. regelmäßige Reports zu KI-Traffic, Eskalationswege bei Problemen mit Googlebot oder Ad-Partnern).

So entsteht schrittweise eine kombinierte Strategie aus:

- Analyse und Monitoring,
- technischer Steuerung,
- vertraglichen und regulatorischen Hebeln,
- sowie klaren internen Regeln.

Genau diese Kombination wird darüber entscheiden, wie viel Gestaltungsmacht Publisher im KI-Zeitalter über ihre Inhalte behalten und ob KI-Bots am Ende nur eine Bedrohung oder auch eine neue Erlösquelle darstellen.

Nur wenn Publisher Transparenz, technische Steuerung und klare interne Prozesse zusammenführen, können sie ihre Inhalte im KI-Zeitalter wirksam managen. Gleichzeitig entsteht Raum für neue Modelle, in denen KI-Zugriffe fair und nutzungsbasiert monetarisiert werden. So entsteht eine strategische Chance, digitale Wertschöpfung neu zu gestalten.

Bundesverband Digitale Wirtschaft (BVDW) e.V.

Der Bundesverband Digitale Wirtschaft (BVDW) e. V. ist die Interessenvertretung für Unternehmen, die digitale Geschäftsmodelle betreiben oder deren Wertschöpfung auf dem Einsatz digitaler Technologien beruht. Mit seinen Mitgliedern aus der gesamten Digitalen Wirtschaft gestaltet der BVDW bereits heute die Zukunft – durch kreative Lösungen und modernste Technologien. Als Impulsgeber, Wegweiser und Beschleuniger digitaler Geschäftsmodelle setzt der Verband auf faire und klare Regeln und tritt für innovationsfreundliche Rahmenbedingungen ein. Dabei hat der BVDW immer Wirtschaft, Gesellschaft und Umwelt im Blick. Neben der DMEXCO, der führenden Fachmesse für Digitales Marketing und Technologien, und dem Deutschen Digital Award richtet der BVDW auch den CDR-Award, die erste Preisverleihung im DACH-Raum für Digitale Nachhaltigkeit und Verantwortung sowie eine Vielzahl von Fachveranstaltungen aus.

OVK

Der Online-Vermarkterkreis (OVK) im BVDW ist die Interessenvertretung der Online-Display- und -Video-Vermarkter am deutschen Werbemarkt. Er setzt sich für die Stärkung des nationalen Online-Werbemarktes und die Erhaltung seiner Angebotsvielfalt ein. Gemeinsam mit den Marktpartnern entwickelt und fördert er Standards und Regelwerke. Als OVK liefert er mit seinen Mitgliedern Orientierung und stellt Markttransparenz her. Er agiert lösungsorientiert; Qualität, Nachhaltigkeit und Zukunftsfähigkeit stehen im Mittelpunkt der Arbeit.

Kontakt

Nicole Dreyer, Senior Programm Managerin, dreyer@bvdw.org

Bundesverband Digitale Wirtschaft (BVDW) e.V.
Obentrautstraße 55, 10963 Berlin
www.bvdw.org