AI SEO In Detail

GA4 bot-filter checklist — is your traffic real?

Episode 15 · Filtering bots in GA4
youtube.com/@AISEOInDetail

Do them in this order: hostname first, bot patterns in clusters, then GA4 against Search Console. No single sign is proof.

1 Filter one: hostname — before any other analysis

- ☐ Break **sessions by hostname** (the domain the visit was recorded on). A healthy property shows only your own live domain (and its www version).
- ☐ Keep only hostnames that are **yours and live**. Old domain, test server, translation proxy, a domain nobody recognises: all noise.
- ☐ Put the filter in **every report and every API query** — ideally inside the one shared function all your scripts use.
- ☐ **Before you reuse a property** (domain change), ask it one question: sessions by hostname. A property accepts data from anywhere that sends it its ID.

The case: a property reused after a domain move showed 53,574 sessions on the retired domain and 55 on the current one over about five weeks — 99.9 % noise, and the dashboard added it all together.

2 Bots that get past the filter

GA4 already excludes **known** bots automatically. You can't switch that off or see what it removed. Look for patterns, **two or three at once in the same segment**:

- ☐ **Zero engagement** — no engagement time, one page, gone; in a neat block of thousands.
- ☐ **One place** — a sudden group from one city or country that isn't your audience (often a data center).
- ☐ **Odd timing** — direct-traffic spikes at 3 a.m., every night, no campaign.
- ☐ **Missing values** — language, screen size or landing page "(not set)". A real browser fills them in.

3 Engagement as a signal

Engaged session = lasts more than 10 s, has a key event, or views 2+ pages. Read engagement rate **and** average engagement time, per segment: real visitors give a spread, bots the same few seconds again and again.

Segment (example site, 90 d)	Engaged	Time
Whole site	~55 %	~45 s
United States (the audience)	~61 %	~50 s
Singapore, Brazil, Hong Kong (700–1,200 sessions each)	Brazil, HK < 10 %	2–3 s
Landing page "(not set)" (~1,200)	2 %	< 3 s
Direct (~15,000)	< 30 %	14 s

4 The Search Console sanity check — never trust one source alone

- ☐ Search Console counts **clicks from Google**; GA4 counts **sessions**. They never match exactly. GA4 is often *lower* (declined cookies, blockers) — that is normal.
- ☐ GA4's "Organic Search" channel counts **every** search engine (Bing, DuckDuckGo, Yahoo…). **Narrow it to source = google** before comparing. Compare weekly, same weeks.
- ☐ A **steady** ratio is a difference in counting. A **drifting** ratio is a question: GA4 can't have counted Google visitors Google never sent.

Week	GSC clicks	GA4 sessions, source = google	Ratio (GA4 ÷ GSC)	Steady / drifting?

Example: all organic = 1.6× the clicks; Google-only = within about 12 % for weeks, then 1.17→1.22→1.34→1.50. The check doesn't give the answer (fake Google referrer? attribution change?) — it tells you where to dig.

5 What to do with bad traffic

1. **Segment it out** of your analysis. Data already in GA4 can't be deleted — the filter must live in your reports and queries.
2. **Write it down:** a short list of known-bad hostnames, sources and patterns, with the date you found them.
3. **Stop the source** where you can: remove the tracking tag from a retired domain, or shut the domain down properly.

Example figures are from an anonymised site; no domains or search terms shown.