CMA Publisher Conduct Requirement: OpenAttribution Response

To: Competition and Markets Authority **Re:** Google Search Conduct Requirements Consultation, Publisher CR **From:** [redacted] OpenAttribution **Date:** February 2026

1. Introduction

This response focuses on crawler separation and the design of attribution and transparency mechanisms.

OpenAttribution develops open-source tools for measuring how content contributes to AI-generated outputs. These tools are early-stage. We reference them below as evidence that the approaches we propose are technically feasible, not to advocate for any specific product.

Summary of key positions: The controls framework is weakened by the circumvention carve-out in Interpretative Note 5 and by the practical costs of exercising opt-outs. Crawler separation deserves reconsideration. Across all seven consultation questions, the cross-cutting issue is that platform self-reporting is insufficient — the CR should require independent verification of attribution and transparency data from the outset, to ensure a reliable evidence base for the 12-month review.

2. Crawler Separation Should Not Be Dismissed

The CMA's provisional assessment (paras 4.82-4.88) concludes that crawler separation is disproportionate relative to improved publisher controls. We disagree.

The cost argument deserves more scrutiny. The CMA estimates third-party costs of £25-50 million annually from increased crawling (para 5.10). Google claimed its own costs would be at least £150 million per year (footnote 104), but the CMA itself notes these estimates "were not made with a high degree of confidence" (footnote 123). Google already operates multiple crawlers. As the CMA acknowledges (para 5.8), a third party with direct technical knowledge considers this technically feasible. Given the CMA's own scepticism about Google's cost figures, and the illustrative benefit range of £32-111 million per year (para 5.52), the CMA should press Google harder on the actual incremental cost before concluding crawler separation is disproportionate.

The controls framework has a structural weakness that crawler separation would solve. Interpretative Note 5 (para 3c of the CR) draws a distinction between active circumvention (eg paying a third party to crawl on Google's behalf) and passive acquisition from pre-existing datasets. The CMA considers it "reasonable for Google to acquire such content through open-source data sets, where these datasets have obtained content legally." The practical effect for publishers is the same regardless. A publisher who blocks Google-Extended can still have their content ingested for AI training via Common Crawl and similar corpora. These datasets contain years of content compiled before publishers had any AI-specific opt-out mechanism, and blocking CCBot now only affects future crawls — not the archive already in circulation. With separate crawlers, a publisher's decision to block AI crawling is enforced at the point of access, removing reliance on downstream controls whose effectiveness depends on how narrowly "circumvention" is defined.

Controls without structural separation leave the compelled consent problem largely intact. The CMA's framework relies on controls, transparency, and attribution working together to shift the bargaining dynamic. Even in combination, these measures depend on publishers being willing to exercise the opt-out. The CMA correctly identifies that publishers have no realistic option but to allow Google's crawler (para 4.6), and its own evidence shows that exercising existing controls carries severe costs: snippet removal reduced traffic by nearly 50% (para 1.11). The proposed CR mitigates the ranking penalty, but the broader dynamic persists: most publishers will not exercise the opt-out if doing so risks any reduction in visibility. Without structural separation, the framework relies on publishers taking an action that market dynamics discourage.

We recommend the CMA reconsider crawler separation, or at minimum tighten the circumvention provision in Interpretative Note 5 to prevent acquisition of opted-out content via third-party datasets. If the CMA proceeds without crawler separation, the attribution and transparency mechanisms must be robust enough to compensate, and independently verifiable.

3. Responses to Consultation Questions

6.2(a): Separate Controls for Training and Grounding (ref. para 4.13)

Separate controls are essential. Training is effectively permanent (content is absorbed into model weights with no practical mechanism for removal). Grounding is per-query and revocable. Attribution is straightforward for grounding but not currently feasible at production scale for training. A publisher may reasonably permit one while refusing the other. A bundled control denies that choice, pushing publishers toward blanket blocking that harms consumers.

Example: a product review site may welcome grounding (cited reviews that drive traffic back) while refusing training (content absorbed into model weights permanently, with no attribution and no revocability). A bundled control forces acceptance of both or neither.

Robots.txt was designed for crawl management and is not fit for purpose as a licensing tool. It cannot express the training/grounding distinction and provides no verification that the directive was honoured. The CR may need to go beyond robots.txt-based controls.

6.2(b): Page-Level Controls for Google-Extended (ref. para 4.25)

Page-level controls are more proportionate than domain-level controls. A publisher may want product reviews available for AI grounding (where attribution drives traffic) but not investigative journalism (where paraphrasing without context could misrepresent the work). Domain-level controls force an all-or-nothing choice that pushes toward blanket blocking.

Page-level controls are only useful if publishers have the data to exercise them intelligently. Without performance data broken down by content type, publishers are guessing. The transparency requirements in 6.3 are a prerequisite. So are the Fair Ranking CR's protections — publishers will not exercise these controls if doing so risks their traditional search ranking.

6.3(a): Per-Feature Performance Data (ref. para 4.45)

Per-feature data should be required. The CMA provisionally considers aggregated data may be more helpful (para 4.44). We disagree: aggregated data obscures precisely the distinctions publishers need. Per-feature breakdowns also risk becoming vanity metrics if they stop at impressions and clicks. A publisher who learns "your content appeared in 5,000 AI Overviews" still cannot determine which content types drive conversions or whether AI traffic is more or less valuable than traditional search traffic.

The CMA should require per-feature data for compliance monitoring (did Google honour the opt-out?) and outcome-based metrics: contribution to downstream clicks, conversions, and engagement, broken down by content type.

6.3(b): Click Quality (ref. para 4.50)

The CMA's second option is the right one: facilitating publishers to calculate click quality themselves, not relying on Google's self-reported metrics.

Google claims AI Overviews send "higher quality traffic." Publishers cannot verify this. Google has an inherent incentive to define "quality" in whichever way favours its AI features. The digital advertising industry faced this exact problem for years — trust only emerged when independent verification became possible (see Section 4).

At minimum, Google should ensure referrals from AI features are identifiable via HTTP referrer headers or URL parameters, distinguishing AI Overview clicks, AI Mode clicks, and traditional search clicks. Publishers can then measure outcomes using their own analytics.

6.4(a): Blocking-Reason Mechanism (ref. para 4.67)

Yes. Aggregated blocking reasons would surface systemic patterns. If most publishers cite insufficient attribution, the attribution requirements need strengthening. If most cite traffic cannibalisation, the transparency metrics are not demonstrating value. A structured taxonomy (eg `insufficient_attribution`, `traffic_cannibalisation`, `no_compensation`, `misrepresentation`, `brand_damage`) gives Google actionable feedback and gives the CMA an evidence base for the 12-month review.

A dropdown taxonomy with optional free-text in Search Console is minimal burden on publishers and negligible cost to Google.

6.4(b): Attribution Metrics and Dissemination

Citation counting is insufficient. Microsoft's AI Performance feature in Bing Webmaster Tools (launched 10 February 2026) shows publishers how their content is cited across Copilot and Bing AI summaries. This demonstrates that providing AI attribution data to publishers is feasible for a major search platform. Requiring equivalent data from Google is clearly proportionate. Even Microsoft's tool currently tracks citation frequency only, without click-through data — illustrating that citation counting alone is an incomplete picture.

Citation counts do not tell publishers whether their content was the primary source or a footnote, how credit is split across multiple sources, or whether content was cited approvingly or to contradict. The CMA should push toward contribution scoring: measuring not just that content was used, but how much it influenced the outcome. We have built a prototype demonstrating this is technically feasible, though not yet deployed at production scale.

The critical question is who computes these metrics. If Google computes and reports its own attribution scores, publishers face the same self-reporting problem discussed in Section 4. The CMA should require Google to provide raw event data (retrieval, citation, display, engagement) via Search Console and a standardised API, enabling independent verification.

6.4(c): Consumer Benefits and Trust

The Publisher CR can deliver consumer benefits, but only if attribution is independently verifiable, not just present. Consumers have no way to assess whether an AI-generated response is based on reputable sources, consulted multiple perspectives, or over-indexed on commercially motivated content.

At minimum, consumers should see which sources informed each AI response, with clickable links. This also matters for brands: as AI responses increasingly influence purchase decisions, brands need to know whether their content investment is fairly represented.

4. Independent Verification: The Cross-Cutting Issue

Across all seven questions, one theme recurs: platform self-reporting is insufficient. The CMA should not create a regime where compliance is self-certified.

The proposed transparency and attribution mechanisms are designed to close the information asymmetry between Google and publishers. But if Google both generates the data and reports it, the asymmetry persists in a different form.

The digital advertising industry learned this. Platforms self-certified viewability, reach, and fraud metrics for years. Trust only emerged when independent verification became possible. The CMA should require independent verification from the outset, either by mandating standardised data formats or by requiring sufficient data access for third parties to verify compliance.

This matters most for the 12-month review. If the evidence base is entirely Google-generated, it will not support the decisions the CMA needs to make.

5. Summary of Recommendations

1. **Reconsider crawler separation**, or tighten the circumvention provision in Interpretative Note 5 to cover acquisition of opted-out content via third-party datasets.
2. **Require independent verification of attribution.** Google should provide raw event data via standardised APIs. If the 12-month review's evidence base is entirely Google-generated, it cannot function as intended.
3. **Require outcome-based metrics alongside per-feature data.** Impressions and clicks are necessary for compliance monitoring but insufficient for publisher decision-making.
4. **Implement a structured blocking-reason taxonomy** in Search Console to build the evidence base for the 12-month review.

We are available to engage further on any of these points.

[redacted]

- [redacted]
- [redacted]