ONE VOICE

Mitigating disinformation, hate speech and objectionable content

Disinformation, misinformation, hate-speech, harmful content, fake news: rising awareness of this type of internet-based content has been brought into focus following the global condemnation of the killing of George Floyd in the US. The subsequent Black Lives Matter protests triggered an upsurge in reactionary and often coordinated extremist content masquerading as fact and designed to whip up societal divisions.

This material isn't just spread through major social media platforms (where it also lands); it's often structured on sophisticated appearing web sites, rapidly established and monetised through the ad tech supply chain providing their site owners ad funding.

And perhaps even more concerning for brand owners is that the juxtaposition of premium brand ads can also provide such sites with the oxygen of credibility. Big brand ads can equal endorsement in the eyes of some, thus reinforcing the voracity of the information they're consuming.

The Global Disinformation Index (GDI), a not-for-profit UK based organisation which aims to disrupt, defund and down-rank disinformation sites, has estimated that some $75 million ad spend ends up on 20K+ disinformation sites worldwide. Its latest report (8 July 2020) has counted 480 English language COVID-19 disinformation sites accruing $25 million in mainstream ad revenue and which spread COVID-19 disinformation such as conspiracy theories and false cure information.

This is a serious and growing brand safety issue – indeed a community safety issue - that needs brands to take urgent steps to manage and control, so they don't appear in such environments.

The nuances involved in managing processes to avoid disinformation and misinformation require a different approach to keyword blocking. ISBA has consulted with three of the main content verification vendors (DoubleVerify, Integral Ad Science and Oracle Data Cloud) to understand how such content is recognised across the billions sites and pages that make up the web, and what active steps need to be taken by brands to protect themselves against the very real threat of appearing adjacent to and funding this kind of content or even on sites broadcasting misinformation purporting to be fact.

We have also consulted with the Conscious Advertising Network (CAN) and the Global Disinformation Index for insight and guidance for ISBA members.

Content classification categories vs keyword blocking

Keyword blocking is the simplest tool in managing brand safety for a 'brand suitable' or safe environment. In fact most content verification vendors, would urge brands to move on from keyword blocking to content category management.

Simple keyword-based exclusion lists are blunt tools, often eliminating valuable inventory and audiences from media plans and potentially depriving premium news sources of critical revenues, news sources that support quality journalism. We covered this extensively at the outset of the COVID-19 outbreak with guidance for brands.

The route to building protection against brand placement against disinformation and inflammatory content is to understand and work with content classification systems that work with content categories.

The principle differentiator is that keyword blocking is an automated process and unintelligent in the way it works to include or exclude inventory, often resulting in excluding the high value inventory brands are actually seeking to invest in.

Content category classification doesn't rely on single term mentions or keyword-type identification. It's the result of a complex blend of machine and human learnings and interventions. Content expertise built out of a blend of linguistic, computer programming and classification skills has allowed complex algorithms to be built which can recognise types of language and contexts to flag risk factors with a site or web page.

Combining AI and human intelligence

Intelligent algorithms can also recognise patterns belonging to the structure of websites built to spread inflammatory and politically or socially divisive content, typically created by activist groups – even illegal ones – seeking to distribute disinformation.

There are characteristic traits that identify such sites. These include URLs appearing overnight purporting to be 'news' sources. Differentiating them from a standard premium news site such as *The Times*, *Guardian*, *Huffington Post* and BBC is that every 'item' on the site is posted at the same time. Premium news sites are regularly refreshed with multiple stories carrying different upload times. These are 'machine readable' differentials.

The CV vendors continue to amass intelligence and ensure their algorithms evolve to capture are set to identify possible risks against ads appearing on such sites and against unsuitable content.

Content classification uses complex algorithms that seek to understand the surrounding content and its context by examining the whole page. The CV vendors, such as DoubleVerify, Integral Ad Science and Oracle Data Cloud with whom ISBA consulted, all agreed that keyword blocking without context and a deeper understanding of content classification, does not provide brands with effective brand safety and brand suitability protection.

Each of the CV vendors, has built content classification categories, and brands and their agencies can work with them to create inclusion and exclusion lists suitable for individual brands. For instance, this example of hate and disinformation content categories that brands can use to create exclusion lists, uses classifications that DoubleVerify has built and maintains. It illustrates the nuances required to target brand safe and suitable destinations for ads:

Category	Classification Description
Inflammatory News and Politics	News or political content associated with or exhibiting inflammatory points of view; potentially fake, unreliable or unsubstantiated information; significant political intolerance, hateful or threatening rhetoric; or other significantly controversial elements.
Hate Speech	Content containing, or about, biased or derogatory language or behaviours towards an individual or group based on their race, nationality, gender, sexual

	orientation, religion or disability. Examples include articles about race-related controversies or violence, hate crimes, the KKK, holocaust or holocaust deniers, as well as vicious or derogatory language in article comments.
Hate/Profanity	DV's Hate Speech category includes content related to hate speech, including common terms and groups. DV identifies hate speech terms and groups through ongoing research and reviews of trusted public and academic resources—keeping up to date with any new trends and developments. In addition, DV utilises a written, internal policy when assessing whether or not content should be considered as hate speech. The policy takes into account a number of factors, including whether speech is targeted at a protected class. Hate/Profanity: Cyberbullying Content that harasses, threatens, humiliates or hassles other people including web pages, videos, images or profiles on social-networking sites making fun of others.

This is an example from a single vendor only and it is vital that advertisers work closely with their own agencies and CV Vendors to understand the various content classification categories and segments that can be subscribed to, how they are built and maintained and, indeed, whether custom categories can be created.

The risks:

In short – TRUST. Advertising within or against disinformation, hate speech, misinformation, inflammatory environments and all other forms of objectionable content, is bad for a brand's reputation. And that is a direct hit to the bottom line, impacting immediate sales and future marketing costs to recover.

In a Harris Poll commissioned by DoubleVerify, 2/3 of consumers said they would be likely to stop using the brand or product if they viewed the brand's digital ad next to false, objectionable or inflammatory content.

That same research revealed that 87% of people feel that brands bear the responsibility for ensuring their ads run adjacent to content that is safe.

Integral Ad Science (IAS) points out that 72% of CMOs face pressure to secure brand trust and gain tighter control over their reputation.

Currently, and sadly, advertising is sitting at the bottom of the Edelman Trust Barometer. Addressing where your advertising appears, is an important step in mitigating reputational risk, both individually as an advertiser and representing advertisers as a whole.

How brands should consider addressing mitigating risk against unsuitable content:

1. Keyword blocking should not be used to disqualify content environments.
2. Standard and custom brand safety categories will mitigate brand's risks without over-excluding.
3. Brands should only use technology that can evaluate the entire page — not just the URL.
4. Brands need to understand their risk tolerance, balancing risk and opportunity based on their individual needs. What is considered safe for one may not be for another, especially as we start to see sub-theme content arise from these movements (i.e. standing in solidarity, positive/peaceful protests, etc). To what extent are they risk averse or risk tolerant?
5. Brands should seek to create custom brand safety segments according to their own brand voice and unique consumer insights.
6. This is particularly important for brands that wish to actively appear against positive coverage of social movements.
7. **Critically, every brand should sit down with their agencies and other intermediaries and ensure that urgent steps are taken to establish content classification categories are embedded into their media planning and transaction processes.**
8. As a first step brands can look at current processes and corporate guidance and measure them against three of the six CAN manifestos: Hate Speech, Fake news and Disinformation and Diversity and Inclusion.

And finally…

Words matter – at their best they can inspire better actions - at their worst they can cause harm and incite violence. You may have noticed in this piece, the use of the terms 'inclusion lists' and 'exclusion lists' when the expectation may have been to read 'black or white'.

In line with the recent advice the WFA (World Federation of Advertisers) issued through its Global Alliance for Responsible Media (GARM), ISBA is encouraging all members to cease the use of the racially-loaded terms *blacklist and whitelist* when referring to the content and sites they want excluded or included in their media campaigns.

Our request is simple:

1. **Change the narrative:** Refer to content and sites you want included in your campaigns as ***Inclusion Lists***, and conversely content and sites you want eliminated in your campaigns as ***Exclusion Lists***.

2. **Activate your network:** As marketing leaders and representatives of your organisations, please cascade this shift and politely call out the need for change when you see others not doing it.

3. **Embed it in your operations:** Update all operating documents and especially contracts between marketers, agencies, platforms and technology partners to reflect this shift.

Thank you for supporting this shift.

More resources to help you bone up on managing where your ads land

- Brand Safety Whitepaper – best practices for keeping your brand in suitable environments across the web. Supplied by Oracle Data Cloud

- Why Context Matters Now More Than Ever For Brand Safety – resources and information for advertisers for understanding the nuance of brand suitability. Supplied by Oracle Data Cloud
- Content classification – an overview and approach by DoubleVerify
- Tackling hate speech and disinformation: resources for advertisers by Integral Ad Science (IAS)
- The History of Fake News – DoubleVerify video
- GDI's July 2020 report: Ad funded COVID-19 Disinformation: Money, Brands and Tech
- The Content Verification Guide – best practice guidelines from the AOP, IAB UK, IPA, ISBA, JICWEBs and Newsworks
- Conscious Advertising Network's manifesto on Hate Speech
- Conscious Advertising Network's manifesto on Fake news and disinformation
- Conscious Advertising Network's manifesto on Diversity and Inclusion

Written by Clare O'Brien, Head of Media Effectiveness and Performance

ISBA Media Team

July 2020