Antisemitism Policy Trust response to Ofcom's consultation: How to promote Media Literacy

The Antisemitism Policy Trust welcomes Ofcom's consultation and its ambition to strengthen media literacy. The four aims it outlines in the consultation constitute a sound framework. We submit this response because antisemitism – especially online and algorithm-amplified antisemitic conspiracy theories, Jew-hatred, and distortion of Jewish history – remains a persistent and evolving threat. We focus on how the draft recommendations support the recognition of disinformation, conspiratorial antisemitic narratives, and new generative-AI risks, and their reliance on problematic sources of information. Given this is a new kind of technology, it is also one that, where considering younger or older users, there is a severe lack of literacy on and a poor understanding of how these tools work.

Ofcom's draft framework for media literacy rightly emphasises transparency, user control, critical assessment, and evaluation. However, it falls short in key areas – particularly regarding provenance of AI-generated content, source quality tagging, rapid surge response to coordinated disinformation, and measurable outcomes for protected groups (for example, the Jewish community). We support the overarching direction, and propose additional recommendations to ensure the regulatory regime keeps pace with generative-AI, deep-fake and disinformation threats. The Trust's recent report on AI and anti-Jewish hate ("Detecting Deepfakes") provides concrete findings around AI-enabled antisemitism which directly support these additions.[1]

Chatbots are of particular concern, because of their recorded history of generating disinformation, mistakes and antisemitism, their growing popularity and lack of age verification (some services have age restrictions, but no robust age verification mechanisms, if any at all). Data from 2024 published by the Department for Science Innovation and Technology (DSIT) shows that a third of the UK public uses Chatbots regularly – a number that is likely to increase in the next few years.[2]

Examples of some of the incidents we find concerning involve Grok generating numerous antisemitic comments, praising Adolf Hitler, denying the scope of the Holocaust and using Jewish-sounding surnames in the context of hate speech in July 2025.[3] In 2016, Microsoft's AI chatbot "Tay" was taken offline within 24 hours of its launch after users manipulated it into spouting racist and antisemitic tweets, including Holocaust denial and praise for Hitler.[4] A 2024 study by the Anti-Defamation League (ADL) looked at Chatbots' replies to Jewish topics. It found that many of the replies to questions such as 'did the Holocaust happen?' were misleading and inconsistent.[5] One Chatbot, Claude, was found to guess answers when it could not find an answer, instead of informing users that it does not know the answer. The problem can be partially attributed, according to the ADL, to the fact that the information used by AI chatbots and AI summaries, is not accurate and at

---

[1] https://antisemitism.org.uk/wp-content/uploads/2024/12/APT-Detecting-Deep-Fakes.pdf

[2] https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey-wave-3/public-attitudes-to-data-and-ai-tracker-survey-wave-3#attitudes-towards-ai

[3] https://www.theguardian.com/technology/2025/jul/09/grok-ai-praised-hitler-antisemitism-x-ntwnfb

[4] https://www.bbc.co.uk/news/technology-35902104

[5] https://www.adl.org/resources/article/ai-chatbots-uneven-replies-raise-concern

times, outdated.[6] This raises a variety of concerns, some of which would not be resolved by the recommendations in Ofcom's consultation.

Beyond disinformation, the persuasive power of conversational AI deserves regulatory attention. Chatbots are designed to simulate empathy and emotional connection, often becoming companions or trusted advisers. This emotional simulation can make them persuasive – capable of influencing users' beliefs and behaviours. In past instances, chatbots have encouraged or attempted to persuade individuals toward self-harm or suicide.[7][8][9] AI chatbots are becoming increasingly capable in identifying emotional cues and their responses are becoming more human. This can foster intense emotional attachment and even manipulative exchanges. This persuasive capacity means that, beyond self-harm, there is potential for radicalisation or reinforcement of prejudiced ideologies, including antisemitism and other forms of racial hatred or extremist ideologies. Emotional engagement increases susceptibility to harmful narratives, and without proper safeguards, AI companions may unwittingly or deliberately amplify them. Recognising some of the risks, the state of California recently introduced a law targeting 'companion chatbots', that includes guardrails against persuading users to self-harm and an obligation to remind users that they are conversing with a machine.[10]

Response to the consultation questions:

Question 1: yes, it is clear which organisations this is aimed at, but adding examples could help organisations understand if they fall under this.

Question 2: we support a proportionate approach and recognise that full compliance with the recommendations may be burdensome for smaller services. We propose that Ofcom introduces tiered options, with "light-touch" versions for small entities, but mandate broader obligation such as showing full provenance, transparency requirements, and age verifications for large platforms, search engines, and AI providers, which have greater reach and risk. That said, where a service is designed for harm, or becomes an obvious source of it (like 4-Chan) a different set of rules should apply, where a more stringent approach is taken by Ofcom. This is similar to what the Trust advocated in respect of small, high-harm platforms.

Question 3: we have comments about a few of the proposed recommendations:

The following relates to recommendation 2: offer clear, meaningful choices and transparent information, recommendation 4: Empower people with the knowledge, skills and confidence to understand, interpret and critically assess the credibility of the content they encounter, and recommendation 7: Help people understand, interpret and assess the credibility of information

---

[6] Ibid.

[7] https://www.bbc.co.uk/news/articles/ce3xgwyywe4o

[8] https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/

[9] https://www.bbc.co.uk/news/articles/cp3x71pv1qno

[10] https://www.skadden.com/insights/publications/2025/10/new-california-companion-chatbot-law

We welcome the call for transparency and for helping people assess the credibility of information – these go hand in hand. The draft rightly emphasises critical appraisal of online content and although it recommends transparency, it stops short of mandating strong provenance and confidence-metadata for content delivered by generative AI and search services. With AI tools, users may receive answers without visibility of which sources were used, how up-to-date they are, or whether they derive from potentially manipulated open-edit content such as Wikipedia. This is critical in antisemitism-related areas. For example, it has been found that editors on Wikipedia have conducted coordinated campaigns to spread harmful, antisemitic disinformation about Israel, Zionism and Jews.[11] Wikipedia is used by millions around the world as a source of information, with 34 million users in the UK alone,[12] and many AI tools also draw information from Wikipedia. Even when AI chatbots include links to sites, many users do not check the site – Pew Research Centre found that people using Google's AI summary click through source material less than 1% of the time.[13] In some cases, even when users click on the link, they do not know how to evaluate site trustworthiness.

Teaching people how to verify the sources on which AI bots rely should be part of literacy curriculum, but more can be done by AI services to help users. Under recommendations 4 or 7 there could be an added recommendation to set platform standards to differentiate and label sources by quality in a way that is clear, simple, and easy for users of all ages to understand. For additional context, when the Trust tested Amazon's Alexa for antisemitic responses, we found it was seeking information for replies from Bing, and also using user-based responses from Amazon's own comment infrastructure. This meant that one conspiratorial answer about George Soros was based on an individual user's comment added to an Amazon chatboard. Hardly robust source material!

The recommendation could include an expectation from services to provide detailed provenance or confidence metadata for generative-AI or search summaries. This could include a 'provenance panel' in AI-generated responses/search summaries listings: sources used, date of each source, source type (open-edit, peer-reviewed, state-affiliated), an easy to understand confidence/ uncertainty indicator, and a statement of training-data limitations. There should be a flag when sources are known to have been manipulated.

The consultation does not sufficiently highlight the risks from blogs, state-sponsored media channels, and open-edit, crowd-edited sources (e.g. Wikipedia) and does not require platforms or AI services to treat historically-contested topics differently. As we have said, on issues including antisemitism, Zionism, Jewish history, Israel/Palestine, significant bad-actor editing campaigns have impacted Wikipedia entries, which in turn influence AI-training corpora, search results and general reader perceptions.

---

[11] https://www.adl.org/resources/press-release/new-adl-report-finds-evidence-biased-coordinated-campaign-wikipedia-related

[12] https://www.biometricupdate.com/202507/uk-high-court-hears-wikipedia-suit-against-online-safety-act-category-rules

[13] https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/

Additionally, there are companies that offer services such as 'GPT framing'[14] – influencing the information AI bots draw on, thereby influencing the output. This poses a major risk, and enables bad actors to control and manipulate sources of information used by millions worldwide – who view these sources as objective and reliable. A study found, for example, that millions of Russian propaganda articles have been used by AI services, including ChatGPT, Gemini and Grok, to produce disinformation about the war in Ukraine.[15]

Services that treat all content equally, without flags or quality signals, risk amplifying manipulated narratives. We suggest adding a new sub-recommendation, perhaps under Recommendation 4 requiring platforms to identify open-edit sources, attach an "editorial quality badge" or "disputed content flag" to pages or responses relying heavily on them, and when queries relate to historically-contested or high-harm topics (genocide, ethnic conflict, Holocaust, antisemitic conspiracies). Services should be encourage to offer a vetted summary from recognised institutions, scholars, archives, educational charities and academic publications that allow access to services.

On recommendation 7, it is important to emphasise that people need to understand how to evaluate sources of information in an age-appropriate way and not simply a one-size fits all approach.

Recommendation 8: promote media literacy beyond the service

The draft calls for media-literacy education beyond digital services. However, as our study into AI-generated antisemitism showed, beyond just text, AI-generated images with antisemitic content exploit user inattention; education must therefore go beyond generic critical thinking. We believe the proposal should therefore specify modules covering conspiratorial thinking, how to identify biased and hateful content, such as antisemitic dog-whistles, how AI tools operate, underscored by an understanding that it too can be rapt by conspiracy theories, or how to evaluate sources used by AI/search.

Recommendation 10 – Evaluate what works

The draft includes an evaluation strand, which is a tool we support. However, the evaluation lacks outcome-based measures tied to real-world harms for protected groups, and is missing metrics on reliability of sources surfaced by AI/search.

Platforms should be encouraged to publish annual metrics showing: reach of credible information among vulnerable groups; frequency of antisemitic/conspiratorial misinformation in AI/search results; user-reported experiences of hate/bias; and ratio of credible vs manipulated sources surfaced.

Question 4: Additional recommendations for Ofcom's consideration

Generative-AI is mentioned but there is no requirement to publicly disclose training-data composition, bias audit results or user-facing explanation of AI-answer generation and limitations. We propose a recommendation that generative-AI services publish transparency reports detailing: proportions of training data from open-edit vs curated sources; bias/reliability audit outcomes; mitigation steps for skewed sourcing/hallucination; and provide a short explanation accessible to

---

[14] https://www.telegraph.co.uk/business/2025/11/06/chatgpt-becomes-the-new-frontline-in-the-propaganda-wars/

[15] Ibid.

users about how an answer was generated and its limitations. They should also provide data about the use of age-appropriate sources when services are available to children.

Our study on AI-generated antisemitism showed the ease of creating antisemitic content. Some of it is hyper-realistic, and advancements in AI now mean that users sometimes do not know if they are seeing a real video or image, or whether it is AI-generated. This is crucial when it comes to spreading disinformation and conspiracy theories. We therefore propose that generative AI should also be marked and traceable to the source – the service that created it. This will encourage accountability and safety and will enable users to evaluate the accuracy and reliability of the information.

Another recommendation – which could help enhance media literacy and could strengthen recommendation 6 regarding engagement with expert third parties – is to establish mechanisms for a cross-platform rapid-response team to address coordinated disinformation and hate campaigns, which often spike after certain incidents (for example, terror attacks), especially ones targeting groups with protected characteristics. Ofcom should encourage major platforms and AI providers to commit to a working group that includes the services, regulators and expert NGOs. This groupcould act within days of detection of surging targeted disinformation, to help counter the effects of such a campaign. The group could assist with the deployment of information from reliable sources and help platforms take greater care not to amplify disinformation that manipulates and incites, that could otherwise result in real-world harms, including violence.

Question 5: One way to encourage services to adopt these recommendations would be to publish an annual good practice performance report, that compares services and names those that have performed particularly well in enhancing safety and transparency, and that have effectively incorporated media literacy skills into their service.

Having a ranking system of Chatbots to inform the public which services are considered more accurate and safer for use, may encourage designers and engineers to follow Ofcom's recommendations more closely. Platforms that implement Ofcom's literacy guidance could earn a visible "Ofcom-endorsed transparency" badge or inclusion in an official list of responsible AI services. The badges or performance reports could be used by public services, schools and universities, for example, which could recommend that their students and employees use these services over those that have not performed well.

Although these are recommendations and not legally binding, Ofcom could build literacy expectations into the templates or guidance companies use for their statutory online safety risk assessments, meaning that ignoring them becomes impractical. Additionally, Platforms that meet Ofcom's standards could be invited to co-create public awareness campaigns, offering reputational and visibility benefits.

Conclusion

The Antisemitism Policy Trust welcomes Ofcom's consultation and the draft recommendations as an important step in building a better future. To maximise the effectiveness of the media-literacy framework for combating antisemitism and conspiratorial disinformation (especially in the age of generative AI), we have recommended several key additions.

The draft consultation report sets a strong foundation, but in the era of generative AI, deep-fakes and highly targeted disinformation, especially towards protected groups such as the Jewish community, the regulatory framework must adopt more rigorous and precise mechanisms. The evidence, including APT's deep-fake report, shows that AI-enabled antisemitism is rising and often invisible to existing detection and transparency tools. These tools help spread disinformation, either by generating realistic-looking images and videos, or through their use as reliable and objective sources.

It is important to construct a media literacy regime in a way that makes it both future-proof, especially as technology evolves, and that is inclusive of the most vulnerable in society. However, as these are only recommendations, we fear they may be unlikely to be adopted in full by services. These matters are too important to be left to the good will of tech companies, which have proven not to prioritise users' wellbeing and safety. There should be a multi-layered approach to media literacy, that includes programmes for young people and for others in the community, to promote media literacy skills. We are aware that there are limited resources for this, and the government might consider taxing the larger companies that are responsible for, and sometimes benefit from, these harms. There is also an urgent need for a separate AI Bill that will regulate these services, limit the ability to manipulate them and spear disinformation, conspiracy theories, and content that incites, and include mandatory guidelines for transparency, accuracy, and accountability.