CHARITIES AGAINST HATE

Social Media Product Recommendations

Contents

Who we are

The Charities Against Hate collective is a group of more than 40 charities working together for real and meaningful change to ensure social media platforms address online hate which is causing real harm to real people.

About these recommendations

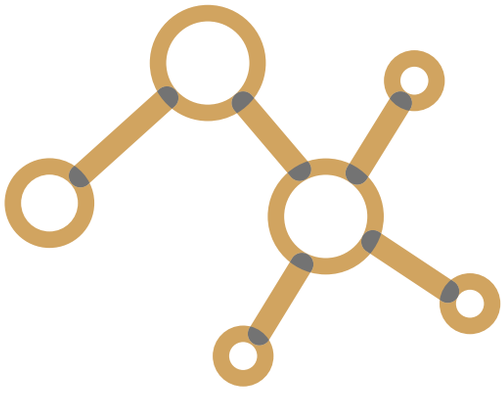
The recommendations suggested here are not exhaustive, but they offer a starting point.

Generally, social media companies have significant resources, both financial and in their employee base. We hope this leads to a better understanding by social media companies that society wants them to put more of those resources into doing the hard work of transforming the potential of the largest communication platform in human history into a force for good.

Accountability

1. Improve existing infrastructure

Social media platforms/ channels should work to develop and improve their existing systems infrastructure to identify, tackle and remove content or activity that harms individual users.



2. Evaluate for discrimination

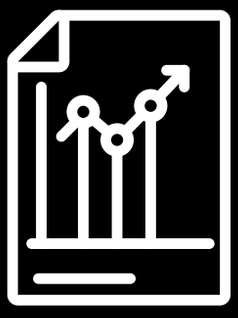
Where practical, the social media platform should have, or appoint a representative at a senior or executive level with the skills, knowledge and expertise to evaluate products and policies for discrimination, bias, and hate.

This person(s) would ensure that the design of product and services and the decisions taken by the social platform fully considers the impact on all communities who use and access the platform.

Further, the potential for online harm, hate speech and discrimination should be considered and evaluated with appropriate timely action taken as required.

3. Conduct independent audits

The social media company should submit to and engage constructively with frequent independent audits which are carried out either by, or with a third party.



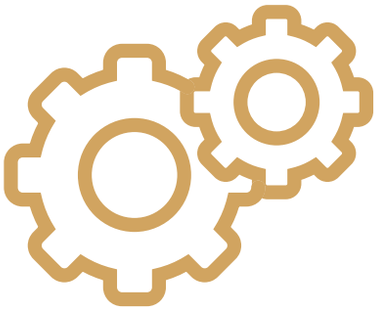
The audit should cover multiple factors, including, but not limited to: online harm, discrimination, hate speech, mis-information and dis-information. Summary results should then be published in a timely way through a publicly accessible website along with the scope of the audit and any other relevant information.

4. Audit removed content

Audits of content that has been removed for policy or terms of service violations should take place.

Where an audit finds that advertising has been erroneously placed with the removed content, an advertiser should be informed and appropriate remedial action taken by the social media platform. This action could be agreed between the advertiser and the social media platform.

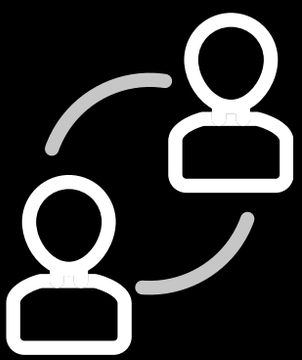
5. Collaborate with charities



Social media platforms should collaborate with charity sector and civil society-based organisations in the UK who work with under-represented groups that are often targets of hate speech and discrimination.

This partnership would be a chance for a range of organisations to identify and flag up instances of potential hate speech, misinformation, distressing trends/challenges and online harm.

6. Work across platforms



Social media platforms should collaborate with each other on tackling all forms of damaging content. Making the internet safer and open to all is in the interest of everyone and should come before competition.

7. Commit long-term

Social media platforms should recognise that this is an ongoing process. This is not an issue that can be 'fixed' once and for all.

By agreeing to tackle hate speech, online harm, discrimination and abuse, social media platforms need to recognise that this is a long-term commitment to make their channels safer and better for everyone who uses them.

Decency

8. Remove hate groups

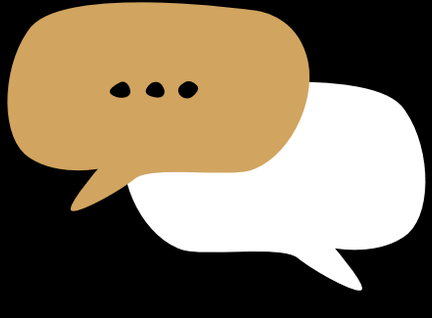
Find and remove social media groups and communities, whether public or private, which focus, for example: on white supremacy, terrorism, antisemitism, violent conspiracies, Holocaust denial, vaccine misinformation, and climate denialism. This list is not exhaustive and is open to amendment.

9. Stricter penalties for individuals

Users who engage in hate speech or targeted harassment of any kind should be subject to stricter penalties and lifetime bans should be considered at a far earlier stage than they currently are.

10. Use informed research

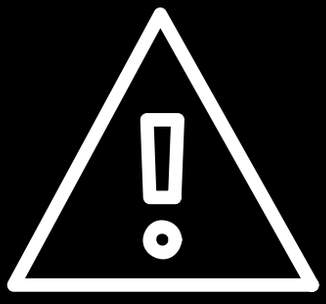
Implement policy changes that are informed by research with relevant target / underrepresented groups to help reduce online hate and harm, the spread of mis/ disinformation, and hate speech on the platform.



11. Prevent the amplification of hateful content

Stop the recommendation and amplification of social media groups and communities with content that, for example, is harmful, discriminatory or associated with hate, mis/ disinformation and violent conspiracies to people who use the social media platform.

12. Develop systems to automatically flag content



Develop appropriate internal systems and tools to automatically flag content based on the factors listed in point 3 of the Accountability section, in public or private communities or groups for human review.

Public or private groups and online communities are not always small gatherings of friends - they can be of a significant size, featuring hundreds of thousands of people as members.

13. Apply all rules equally

All rules relating to hate speech, misinformation and harmful content should be applied equally and consistently to all users of social media.

Politicians, brands, news outlets and public figures should not be exempt and should be held accountable when it comes to spreading hate and misinformation.

Social media platforms should take timely appropriate and equal action where users act against the stated rules and policies of the social media platform.

14. Apply profanity filters

There should be profanity filters as standard on all platforms. These filters should be reviewed on a regular basis by the social media platform.

Where practical, there should be a mechanism which allows people who use the social media platform to update / add content to a profanity list or filter. There should be transparency about what terms are included in a profanity filter.

Support

15. Establish expert teams

The social media platform should establish teams with the expertise, skills and knowledge to review submissions of content shared through the platform which are intended to cause online harm, hate or distress to a person based on their personal characteristic(s), known or otherwise.

Social media platforms should ensure those staff who are directly employed by or subcontracted to the social media company are given the skills and training to identify and understand the various types of harassment faced by different groups in order to adjudicate claims.

16. Provide access to real people

Enable people who use the social media platform who experience hate and harassment, which is severe or sustained to have access to a live employee. There are very few other industry sectors or companies who do not have established ways for victims of a product to seek help.

Appendix

1. Auditing and content moderation changes

- **Content Auditing**
  This needs to be in person moderation either through a community or staff member.
- **Potential change**
  'I find this offensive' button or mechanism to give feedback directly on the platform.
- **Algorithm changes**
  Automation to moderating content.

2. Practical changes social media platforms could make

- Easier flag option for user community to report hate speech.
- Lifetime bans for perpetrators.
- More sophisticated algorithm to ban key words used in hate speech.
- More accurate ban on flagged language per page.
- Profanity filters being fit for purpose.

3. Employee related issues

- How staff working for social media organisations/ platforms are affected by moderating content that is harmful or hate speech and issues.
- Potential support/ resources for mental health support to employees working in social media organisations offered by the UK charity sector.

www.charitiesagainsthate.com