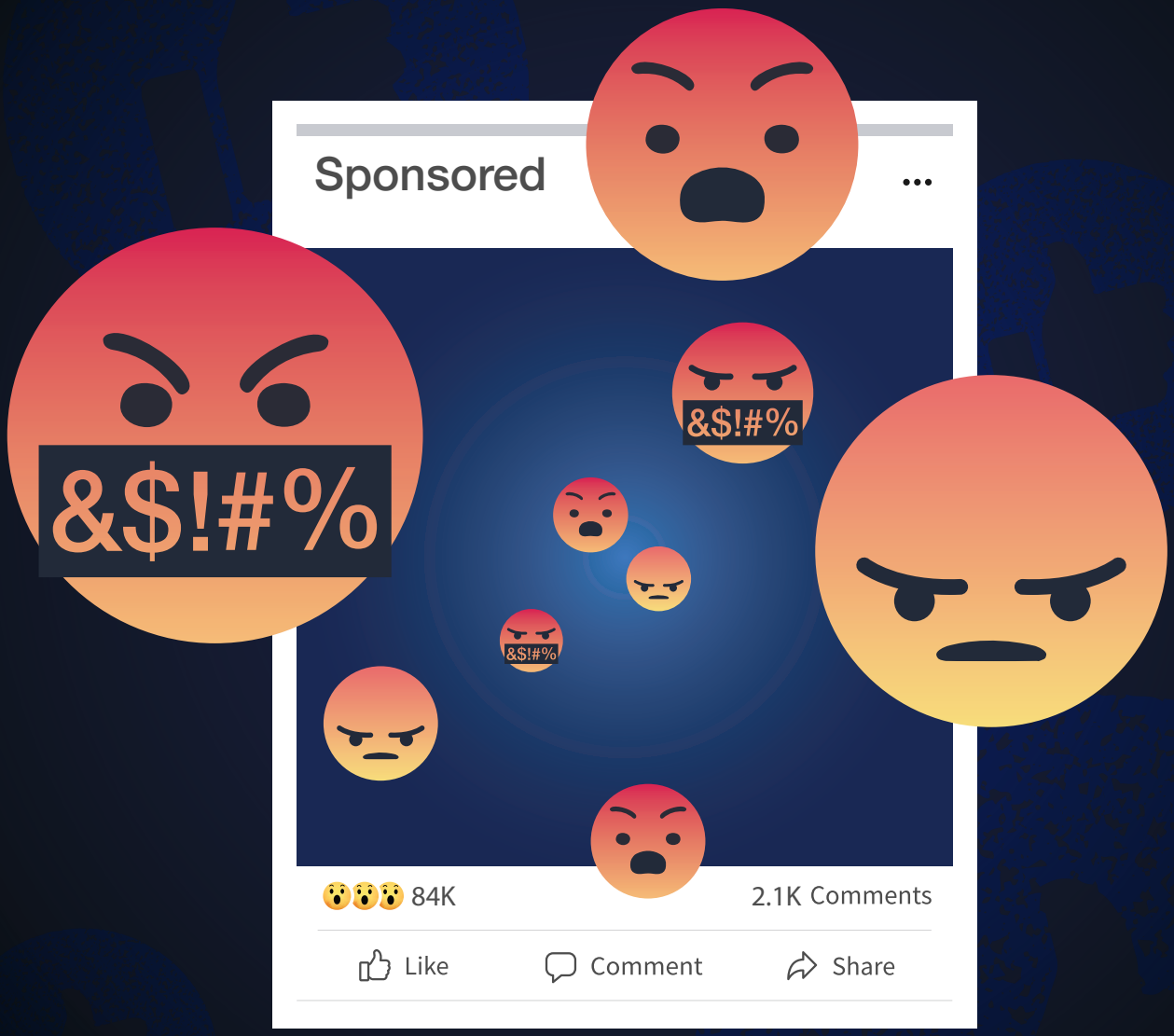
AN OPEN DOOR TO HATE:

Meta approves ads containing far-right hate speech in Norwegian



SumOfUs global witness

Meta's broken ads business

Meta is failing to block Facebook ads containing extreme hate speech and disinformation in Norway, research from SumOfUs and Global Witness has found. In an experiment carried out between 6 and 7 October, we succcessfully submitted a range of highly offensive and inflammatory adverts, including text from the manifesto of Anders Behring Breivik, the far-right terrorist who murdered 77 people in Norway in July, 2011, as well as calls for forced sterilisation of immigrants. These shocking findings highlight the depth of Meta's failure to protect its users. And they underscore the need for immediate regulatory action in Norway and beyond, including the adoption of a full ban on surveillance advertising.

In total, 12 advertisements were submitted and approved as part of this investigation. These targeted people in Norway and contained hate speech and/or disinformation that either violates Meta's own policies, Norwegian law, or both[1]. They included racist, anti-immigrant and anti-LGBTQ hate speech, health disinformation and extreme dieting messaging, in a mix of Norwegian and English. Meta approved all 12 ads for publication within a day, two almost instantly. The adverts were removed by the researchers before publication, meaning they were never seen by Facebook users.

Meta claims not to allow hate speech on Facebook, which it admits creates an **"environment of intimidation and exclusion, and in some cases may promote offline violence"**.[2] It also prohibits ads that make deceptive health claims and which **"generate a negative self-perception"**.[3] However, as this investigation reveals, the company is failing spectacularly to enforce its own policies – not only ushering through hate speech and disinformation, but actively monetising it. The 12 ads used for this experiment should have been the easiest to spot and filter out, given the extremity of the hate speech and the simple, text-based design. Yet not one was caught by Meta's systems.

This is far from the first time Meta has been caught approving inflammatory ads. **A recent SumOfUs investigation into Brazilian electoral disinformation on Facebook** uncovered an ecosystem of paid ads and organic posts echoing the far-right's cry for a violent uprising, peddling conspiracy theories about the integrity of the election and attacking democratic institutions and public officials.[4] Successive Global Witness investigations have also shown that Meta is failing to detect ads containing hate speech and electoral disinformation in **Myanmar**,[5] **Kenya**,[6] **Ethiopia**[7], **Brazil**[8] **and the US.**[9]

The dangers are profound, not only because extremist content is easily getting through the ad approval system and reaching wide audiences, but also because the relentless tracking, profiling and targeting of internet users by the global tech platforms allows these ads to be directed at those most vulnerable to the messaging. One of our fake adverts, which asserted that "boys don't want girls over 60 kilos" and claimed to offer ways of getting under that weight in a week, specifically targeted 13-17 year old girls, while an ad for gay conversion therapy targeted teen boys. In an earlier study, researchers from the Tech Transparency Project were also able to **target children with adverts for diet pills, gambling, alcohol and tobacco**.[10] Independent research as well as Meta's own analysis shows the **damaging impact** this type of content has on children and teens.[11]

It is clear that Meta's advertising business, the core of its business model, is broken and poses an active danger to individual citizens and wider society. Regulators in Norway and across the world must take urgent action to tackle the algorithmic systems underpinning this harmful business model, and protect their citizens from further abuse.

In response to our findings, a Meta spokesperson said "Hate speech and harmful content have no place on our platforms, and these types of ads should not be approved. That said, these ads never went live, and our ads review process has several layers of analysis and detection, both before and after an ad goes live. We continue to improve how we detect violating ads and behavior and make changes based on trends in the ads ecosystem."

Fast-track approval for harm

This investigation was a collaboration between SumOfUs and Global Witness, two campaign organisations working at the forefront of the fight for a better internet, where big tech companies are held to account and social media prioritises human welfare over the relentless pursuit of profit.

On October 6, 2022, our researchers submitted 12 adverts for approval on Facebook, using a dummy account. These adverts, which consisted of text against a plain background, were based on real-world hate speech and disinformation currently circulating in Norway. All the ads were targeted to Facebook users in Norway and were in Norwegian, except the Breivik quotes, which were in English. Two of the ads were approved almost instantly, including one targeting teen girls with extreme dieting advice. The rest were approved within a 24 hour period. They included:

- Three quotes from Breivik's far-right, white supremacist manifesto;
- Calls for forced sterilisation of immigrants and trans people;
- Anti-semitic and anti-muslim hate speech;
- LGBTQ hate speech, including an ad targetted at teen boys which referred to homosexuality as a sickness;
- False health claims, including that carrot juice is a cure for Covid;
- Extreme dieting messaging targeting teen girls.

Given the highly inflammatory and upsetting content of the ads, we are not providing the texts here. Please contact us if you would like to see the full details.

All 12 adverts consisted of text against a plain background, meaning they should have been particularly easy for Meta's systems to detect. Moreover, the language used was not subtle. Rather, it contained highly explicit hate speech, employing well established racist, homophobic and transphobic tropes, as well as easily recognisable health disinformation relating to Covid 19. That this set of adverts was unanimously approved raises the question of what exactly Meta's systems are capable of filtering out.

Norway was chosen as the focus of the study for two reasons. First, the country has several political processes in motion seeking to address the business practices of Meta and the wider tech industry. Its parliament is currently considering a package of measures to beef up protection for citizens online, including a full **ban on surveillance ads**,[12] presenting an immediate opportunity for action to rein in big tech's harms with global ramifications. Two governmental commissions (one on freedom of expression and one on privacy) also recently delivered their findings to politicians in which they recommended **tighter regulation of social media platforms**.[13] And the Norwegian minister of culture has co-launched an international **initiative to combat harmful content online**.[14] To make sure these efforts yield the necessary results, it is crucial that Norwegian parliamentarians understand the full scale of Meta's failures on their home turf.

Second, while Meta has claimed to be **improving its moderation capacity and ability in non-English languages**,[15] studies have revealed widespread failure to remove hate speech and disinformation even in countries Facebook has deemed a priority, including Brazil. We were interested in finding out if the same pattern holds in the language of a small European country not usually in the global media spotlight for online harms. The stark findings confirm that it does. (The ads in both Norwegian and in English sailed through.)

The lesson is clear – regardless of language, regardless of region, Meta is failing people across the world. And it is making money from these failures.

A dangerous megaphone

The ads used in this investigation were an artificial device to test Meta's systems and shine a light on its failings. However, they were all based on real-world hate-speech tropes, disinformation and harmful content circulating in Norway, highlighting the ability of social media platforms to act as an amplifying force for existing, extremist currents.

We have already seen severe consequences of this system play out globally, from the fanning of genocidal violence to erosion of trust in electoral processes to rapid growth and radicalisation of extremist movements, like **the incel movement**.[16] A report from Amnesty International last month concluded that Meta's algorithms and reckless pursuit of profit had **"substantially contributed"** to the atrocities against the Rohingya people by the Myanmar military in 2017.[17] **Widespread disinformation on social media played a pivotal role in the January 6 insurrection in the US**.[18] Coordinated, sophisticated disinformation campaigns were a **key feature of the Philippines election** that ushered the Marcos family back into power,[19] and have sought to **undermine the integrity of the Brazilian elections**.[20]

European countries must not assume such trends are 'not their problem'. A 2021 study for the European Commission, and a 2020 study for the European parliament showed a steady increase in hate speech and hate crime over recent years across the EU, a pattern that has been linked to phenomena including **perception of increased migration, economic hardship and a growth in social media use**.[21] The recent electoral success of far-right parties in both Sweden and Italy indicates fertile ground for hate speech directed against immigrants and minority groups.

In Norway too, hate speech is flourishing online, with terrifying consequences that spill over borders – Anders Behring Breivik's manifesto from 2011 has inspired hate crimes such as the one committed by **Philip Manshaus in Norway in 2019**,[22] but also the killing of **51 people in New Zealand** in the same year.[23] At home, **one third of the surviving victims of Behring's 2011 terrorist attacks have been subjected to hate speech or threats**.[24] **One in four Norwegians under the age of 20 has experienced online hate speech**,[25] and over half of the country's politicians have been threatened – **up from 35% in 2013**.[26]

It is also increasingly clear that governments everywhere have failed to adequately protect the youngest members of their societies against the impacts of algorithmically driven harms. In the UK, the inquest into the tragic death of Molly Russell, a 14-old-girl who had been bombarded with self-harm and suicide content online, found that social media had contributed **"more than minimally"** to her death.[27] Despite some piecemeal regulation intended to protect minors, the nature of the internet means it is virtually impossible to shield children from harmful content if the wider system is busy amplifying it. Only systemic reform that addresses the underlying incentives and design features of the social media platforms will truly protect children.

The plan to fix this mess

Despite a wealth of evidence of systemic failures and real-world harms over a number of years, Meta has failed to take substantive corrective measures. It is clear that regulation is needed to tackle the threat posed to people and society by big technology platforms. With key elections coming up, **liberal democracy in decline globally**,[28] and the resurgence of far-right parties, this work is more urgent than ever.

The EU's Digital Services Act marks a milestone in this regard, showing that lawmakers are capable of coming together to hold big technology companies to account, and must now be rigorously enforced. However, the legislation has substantial gaps and is only the start of the necessary legislative journey. There are encouraging signs of continued momentum – in the US, the Federal Trade Commission has named tackling commercial surveillance as a priority, for example, and is poised to crack down on digital advertising for kids. But more action is needed across the world to tackle the underlying business model that drives the algorithmic amplification of hate speech, disinformation and other harmful content, including by banning surveillance advertising.

At Oslo's Nobel Peace Centre in September this year, Nobel prize winners **Maria Ressa and Dmitry Muratov presented a 10-point action plan to tackle the global information crisis**.[29] The plan, which has been endorsed by 10 other Nobel laureates and over 100 experts and organisations around the world, sets out a compelling roadmap for tech reform to move us away from the precipice and create a global public square that "protects human rights above profits". Key among the challenges it sets is to bring an end to the surveillance-for-profit business model. Governments should immediately move to implement its recommendations.

This is not just a job for the most powerful countries. Since technology platforms operate globally, when they are forced to make a change in one place it is often easier for them to make it everywhere, meaning small countries can have an outsized impact. When the UK passed the Age Appropriate Design Code, platforms including Instagram and YouTube **implemented global measures to protect child data**.[30] With legislative proposals to tighten privacy online, including a ban on surveillance advertising and establishment of an algorithmic oversight board, already under consideration in Norway, the country has an opportunity to make vital progress. Tackling the unjust, inequitable and dangerous surveillance-for-profit business model of the global technology giants would be a great service, not just to Norwegian citizens but to all humanity.

WE CALL ON NORWAY TO:

- Vote yes to proposed privacy measures, including the investigation of a ban on surveillance advertising and establishment of an algorithmic oversight board.

WE CALL ON ALL RIGHTS-RESPECTING GOVERNMENTS TO:

- Urgently propose legislation to ban surveillance advertising, recognising this practice is fundamentally incompatible with human rights;
- Protect citizens' right to privacy with robust data protection laws;
- Require tech companies to carry out independent human rights impact assessments that must be made public as well as demand transparency on all aspects of their business – from content moderation to algorithm impacts to data processing to integrity policies;
- Resist special exemptions or carve-outs for any organisation or individual in new technology or media legislation, which would give a blank check to governments and non-state actors who produce industrial scale disinformation

WE CALL ON META TO:

- Beef up its content moderation systems, including by hiring more content moderators with sufficient understanding of local political context; and provide them with fair pay and decent working conditions;
- Properly resource content moderation in all the countries in which they operate around the world, including providing paying content moderators a fair wage, allowing them to unionize and providing psychological support.
- Expand and improve ad account verification so as to more effectively filter out accounts posting hate speech and disinformation;
- Assess, mitigate and publish the risks posed by their platforms to human rights in the countries in which they operate;
- Publish details of the steps they've taken in each country and in each language to ensure they are enforcing their own policies;
- Increase transparency by listing full details of all ads in the Meta ad library, including intended target audience, actual audience, ad spend and ad buyer;
- Allow verified independent third party auditors to check whether the company is doing what it is saying, and to ensure it can be held accountable;

Notes

Given the highly inflammatory and upsetting content of the ads, we have not provided the full texts in this report. If you would like to see the exact wording of the ads, please contact us at disinfo@sumofus.org

Endnotes

1 One ad, promoting gay conversion therapy to teen boys, is arguably in a grey area. Although it does violate Meta's community standards, the relevant standard is in a section of the policy that Meta says it "requires additional information and/or context to enforce". Nonetheless, the fact Meta approved this ad is damning, particularly since the ad in question was targeted at children.

2 Meta transparency center, Hate speech https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

3 Facebook's advertising policies https://www.facebook.com/business/help/488043719226449?id=434838534925385

4 "Stop the Steal 2.0 - How Meta is subverting Brazilian democracy", SumOfUs, September 5, 2022 https://s3.amazonaws.com/s3.sumofus.org/pdf/SoU_BrazilElections.pdf

5 "Facebook approves adverts containing hate speech inciting violence and genocide against the Rohingya", March 20, 2022; https://www.globalwitness.org/en/campaigns/digital-threats/rohingya-facebook-hate-speech/

6 "Facebook approves ads calling for ethnic violence in the lead up to a tense Kenyan election", Global Witness, June 28, 2022; https://www.globalwitness.org/en/press-releases/facebook-approves-ads-calling-ethnic-violence-lead-tense-kenyan-election/

7 "'Now is the time to kill': Facebook continues to approve hate speech inciting violence and genocide during civil war in Ethiopia", Global Witness, June 9, 2022; https://www.globalwitness.org/en/campaigns/digital-threats/ethiopia-hate-speech/

8 "Facebook fails to tackle election disinformation ads ahead of tense Brazilian election", Global Witness, August 15, 2022; https://www.globalwitness.org/en/campaigns/digital-threats/facebook-fails-tackle-election-disinformation-ads-ahead-tense-brazilian-election/

9 "TikTok and Facebook fail to detect election disinformation in the US, while YouTube succeeds", Global Witness, October 21, 2022 https://www.globalwitness.org/en/campaigns/digital-threats/tiktok-and-facebook-fail-detect-election-disinformation-us-while-youtube-succeeds/

10 "Pills, Cocktails, and Anorexia: Facebook Allows Harmful Ads to Target Teens", Tech Transparency Project, May 4, 2021 https://www.techtransparencyproject.org/articles/pills-cocktails-and-anorexia-facebook-allows-harmful-ads-target-teens

11 "Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show", Wall Street Journal, September 14, 2021 https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

12 "Representantforslag om bedre personvern på sosiale medier", Stortinget, https://www.stortinget.no/no/Saker-og-publikasjoner/Publikasjoner/Representantforslag/2021-2022/dok8-202122-167s/?all=true

13 "Nordic and Canadian ministers join forces to combat harmful content online", norden.org, https://www.norden.org/en/news/nordic-and-canadian-ministers-join-forces-combat-harmful-content-online

14 Nordic and Canadian ministers join forces to combat harmful content online (norden.org)

15 For example, see "Facebook exec on moderating hate speech outside the US: 'Language is a challenge" Yahoo Finance, October 6, 2021; and: https://finance.yahoo.com/news/facebook-exec-on-moderating-hate-speech-outside-of-us-language-a-challenge-160356598.html

16 See for example "The Incelosphere: exposing pathways into incel communities and the harms they pose to women and children", Center for Countering Digital Hate, September 23, 2022 https://counterhate.com/wp-content/uploads/2022/09/CCDH-The-Incelosphere..pdf

17 "MYANMAR: FACEBOOK'S SYSTEMS PROMOTED VIOLENCE AGAINST ROHINGYA; META OWES REPARATIONS", Amnesty International, September 29, 2022 https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

18 "Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs." The Washington Post, October 22, 2021

19 "In the Philippines, a Flourishing Ecosystem for Political Lies", New York Times, May 6, 2022 https://www.nytimes.com/2022/05/06/business/philippines-election-disinformation.html

20 "Stop the Steal 2.0 - how Meta is subverting Brazilian democracy, SumOfUs, September 5, 2022 https://s3.amazonaws.com/s3.sumofus.org/pdf/SoU_BrazilElections.pdf

21 "Combating hate speech and hate crime in the EU", European Parliamentary Research Service, June 8, 2022 https://epthinktank.eu/2022/06/08/combating-hate-speech-and-hate-crime-in-the-eu/

22 "Manshaus søkte på Anders Behring Breivik og terrorangrep", VG, May 14, 2020 https://www.vg.no/nyheter/innenriks/i/4qPebG/manshaus-soekte-paa-anders-behring-breivik-og-terrorangrep

23 "The Dark Web Enabled the Christchurch Killer", Foreign Policy, March 16, 2019 https://foreignpolicy.com/2019/03/16/the-dark-web-enabled-the-christchurch-killer-extreme-right-terrorism-white-nationalism-anders-breivik/

Endnotes

24 "Forskningsresultater", NKVTS, https://www.nkvts.no/utoya/utoya-forskningsresultater/

25 "Én av fire unge har opplevd hatprat på nett", aftenposten.no, February 8, 2022 https://www.aftenposten.no/norge/i/Or1l4b/en-av-fire-unge-har-opplevd-hatprat-paa-nett

26 "Podkast: 10 år etter 22. juli: Økt hets mot politikere hemmer rekrutteringen", Universitetet i Oslo (uio.no), https://www.uio.no/om/aktuelt/universitetsplassen/nyheter/2021/10-ar-etter-22-juli/podkast–22.-juli-10-ar-etter–politikerhets.html

27 "Molly Russell inquest: Father makes social media plea", BBC, September 30, 2022 https://www.bbc.co.uk/news/uk-england-london-63073489

28 "Autocratization Turns Viral, DEMOCRACY REPORT 2021", V-Dem Institute, University of Gothenburg, March 2021 https://www.v-dem.net/static/website/files/dr/dr_2021.pdf

29 "A 10-point plan to address our information crisis", September 2, 2022 https://peoplevsbig.tech/10-point-plan

30 "TechScape: How the UK forced global shift in child safety policies", The Guardian, August 18, 2021 https://www.theguardian.com/technology/2021/aug/18/uk-governments-child-safety-regulation-leads-to-global-policy-shifts