MOLLY ROSE FOUNDATION

Meta's rollback of safety protections – why the Government and Ofcom must act

Policy briefing – April 2025

On 7th January, Meta announced a substantial rollback of its content moderation policies and that it would cease automated detection and proactive enforcement in key policy areas.

Although the company has been reluctant to set out additional detail about the impact of these changes, this amounts to a nothing less than a bonfire of its safety protections. It's highly likely these will result in significant preventable harm to children and young adults, and will deepen the suicide, self-harm, depression and other mental health risks to which they are exposed.

Meta's announcement has been widely and rightly criticised, but it also highlights substantial weaknesses in the design and implementation of the Online Safety Act – weaknesses that will allow the company to proceed with its changes largely unchecked.

This briefing presents the results of new representative polling of adults across Great Britain, and it shows the public wants and expects a stronger legislative and regulatory response in the face of significant weakening of safety measures by large social media sites.

Our results show:

- **the public oppose Meta's plans to stop proactive enforcement by an over two to one margin:** among those expressing a view, 71% of adults oppose Meta's plans to cease automatically searching for and removing content that breaks its safety rules.
- **strong support for additional protections for children:** 88% of adults support social media platforms being made to prevent children being exposed to harmful content, even if the platform allows this for adults. Meta has so far declined to offer assurances that children will be ringfenced from the changes being made to its hate and unacceptable speech policies.

- **a clear appetite to strengthen regulatory protections:** 86% of adults would support a requirement on social media platforms to proactively search for harmful content. Meanwhile, over nine in ten (93%) of adults who express an opinion would support a requirement on social media platforms to ensure changes to their policies aren't used to scale back their existing safety measures (in effect a 'no rollback' provision).

- **Strong support for new conduct-based rules:** 87% believe that social media platforms should be made to ensure managers with responsibility for safety meet minimum standards of conduct. Recent decisions made by several social media companies has renewed debate about the need to build and strengthen the provisions in the Online Safety Act, including the value of a shift towards conduct-based regulation.

The likely impact of Meta's changes on children and teens

Meta's decision to cease proactive detection of harmful content is likely to expose users to significant adverse impacts, and it is particularly likely to expose children and young adults to increased harm.

Following initial criticism, Meta clarified that it is making no change to its suicide and self-harm policies and that it will continue to proactively identify and remove content that reaches its community standards.

Regrettably, this is a highly casuistic position that offers little if anything by way of meaningful reassurance. As MRF research has shown, much of the potential harm faced by children and young people comes from being algorithmically recommended content that doesn't meet the threshold to be deemed violative, but that is likely to be harmful when viewed in large amounts.[1]

Ofcom analysis suggests that large amounts of such content, including combinations of posts that normalise or regularise suicide, self-harm, disordered eating and intense thoughts of depression, worthlessness and despair, can be a primary driver of cumulative harm – and result in substantial harmful impacts to the mental health and well-being of young people.[2]

In practice, Meta's operational changes are likely to substantially increase the risk of cumulative harm among children and young adults. While the company claims that its Teen Account feature will prevent young people from being exposed to sensitive content, there is limited detail about what type of content will be proactively detected, nor the efficacy of these measures.

1 Molly Rose Foundation (2023) Preventable yet Pervasive: the scale and nature of harmful content on Instagram, TikTok and Pinterest
2 Ofcom (2024) Draft risk register issued as part of the consultation on the Protection of Children regulatory scheme

The likely impact on vulnerable young adults

Vulnerable young adults are likely to be at particular risk because of Meta's downgrade of its safety commitments – with young people aged 18-25 unable to benefit from the regulatory or operational protections available to children or adolescent teens.

While the Online Safety Act offers some limited protections, in the form of users being able to choose not to be algorithmically recommended certain types of harmful content, these measures are likely to have limited effect at best – with the onus continuing to sit with users rather than companies to protect themselves from harm.[3]

MRF is highly concerned that, as it is currently drafted, the Online Safety Act contains nothing that can prevent Meta or other platforms further watering down its protections to adult users.

While category 1 services - which are likely to include Meta's platforms - will have a duty to enforce their Terms of Service when the Act is fully in force, there are no minimum standards they must uphold. As a result of Meta's policy and operational changes, young adults are therefore likely to face sharply increased risks, including exposure to suicide, self-harm and highly depressive content.

In turn, we are deeply concerned about the impact on vulnerable young adults, including those at heightened risk of suicide, self-harm and disordered eating, and those with diagnosed mental health conditions.

3 MRF is unaware of any research that suggests young adults experiencing significant problems with their mental health will necessarily have the competency or means to cope with the cognitive load of making such decisions. The OSA does not require any user testing or other research prior to these measures being implemented. Some users may also be unwilling to proactively screen out content even if this may be harmful: some may be concerned about missing content which accurately or otherwise they may consider to be beneficial to them; they may continue to seek out harmful content as a maladaptive coping response; and like all online users, high rates of choice inertia should reasonably be expected.

Polling results

Methodology

YouGov surveyed 2,275 adults in February 2025. This includes a sample of 452 parents with at least one child aged 18 or younger. Fieldwork was carried out online. The figures have been weighted and are representative of all GB adults (aged 18+).

1. The public opposes Meta's content policy changes and its rollback of proactive enforcement

Our results show significant opposition among the public to Meta's plans to scale back proactive enforcement of its community standards. Overall, the public oppose Meta's plans by a stronger than two-to-one margin.

Among those who expressed an opinion, 71% opposed its plans to change the way it finds and removes violative content, which will mean that less harmful content will be automatically detected and removed. Over two in five adults who express a view (44%) say they are strongly opposed to Meta's changes.

While these results are perhaps unsurprising, they nevertheless underscore the strength of public concern about the likely implications of Meta's decision, but also broader worries about the impacts of harmful online content.

In making the announcement, it was striking that Meta CEO Mark Zuckerberg situated his decision exclusively in the context US civil and political discourse. Meta's CEO stated that he had made this decision because he believed that the 2024 presidential election marked 'a cultural tipping point toward once again prioritising speech', and he welcomed a 'new era' in which Meta could get 'back to our roots' and 'focus on restoring free expression' rather than trying to reduce harmful, false or offensive content. [4]

In taking a decision that predominantly reflects US political and speech-based calculations, it is perhaps no surprise that Meta's shift has been poorly received by adults in another of its largest markets, the UK.

The results should powerfully remind the UK Government and regulators that the UK takes a very different approach to how we balance fundamental rights, including safety, privacy and free expression. Our polling shows that **four in five adults (80%) think there should be legal minimum safety requirements for online platforms set by the UK Government and regulators. Just one in ten (11%) believe that platforms should be able to set their own minimum safety standards.**

4 Comments made by Mark Zuckerberg to accompany Meta's corporate announcement

2. Children should be protected from harmful content, even if a platform allows this for adults

An overwhelming majority of British adults (88%) support social media platforms being made to prevent children being exposed to harmful content, even if the platform allows this content for adults. Support for such a new duty rises to over nine in ten parents (91%).

Meta's recent announcement has thrust the balance of protections between children and adults back into the spotlight. While the Online Safety Act requires children to benefit from a higher standard of protection than adults, Meta has so far failed to offer assurances that children will be ringfenced from the changes being made to its unacceptable and hate speech policies, for example through its Teen Accounts feature screening out content that was previously deemed violative.

As a result of Meta's changes, posts that would previously have been removed or downranked – for example, describing women as 'property' or LGBTQ+ people as 'mentally ill' – would no longer be deemed violative of either Facebook or Instagram's community standards.

Following Meta's announcement, MRF wrote to Ofcom asking them to clarify whether they intend to clarify these changes should not extend to children, and whether Category 1 services should be expected to use proactive enforcement to prevent children being exposed to harmful posts.[5] In response, Ofcom has declined to publicly set out its position or to provide any relevant assurances.

3. Strong support for legislative and regulatory strengthening, including new baseline protections for users

The results indicate strong public support for a strengthening of the regulatory protections afforded by the Online Safety Act, with overwhelming support for new duties that could effectively prevent a repeat of Meta's safety downgrades.

By a striking sixteen-to-one margin, UK adults would support a requirement that social media platforms ensure changes to their policies aren't used to scale back their existing safety measures. Once don't knows' are excluded, more than nine in ten adults (93%) would support such a 'no rollback' measure.

There is similarly **strong support for platforms being able to change content rules only if they can provide evidence that the changes won't result in harm.** Among those who express a view, we found strong consensus for the introduction of this backstop measure, with 91% voicing support.

As it stands, regulated platforms will have to undertake suitable and sufficient risk assessments to prevent exposure to illegal content, and to ensure children are unable encounter or access harmful or age-inappropriate material. While large platforms will also have to undertake assessments to identify the impact of platform design on the likelihood that adults will encounter harmful content, the results of this assessment will only apply to users who proactively choose to shield out harmful content. This includes material that promotes and glorifies suicide and self-harm.

5 A copy of the letter sent to Ofcom CEO Melanie Dawes on January 27th can be read in full here: https://mollyrosefoundation.org/wp-content/uploads/2025/01/Ofcom-Meta-final-.pdf

4. Platforms should be required to proactively enforce their terms of service

An overwhelming proportion of adults (94%, excluding DKs) believe that social media platforms should be required to proactively search for harmful content rather than solely being able to rely on user reports.

Meta's own data suggests that almost all of the harmful content it currently actions is the result of its own proactive enforcement initiatives. According to our analysis of Meta's regulatory disclosures, more than 90% of its content moderation decisions that relate to suicide and self-harm content stem from at least partially automated means.[6]

Our analysis raises substantial questions about whether large platforms that opt not to use proactive technology can adequately enforce their terms of service. Platforms that currently opt not to use proactive technology to detect violative suicide and self-harm content moderate tiny volumes of material compared to those that do. For example, our analysis found that over an eight-month period, X (which does not use proactive moderation technology) identified 173 times less violative content than TikTok (which does). [7]

Recent research from the Center for Countering Digital Hate has suggested that Meta's changes could lead to a 97% reduction in content enforcement in key areas, including hate speech, bullying and harassment, and could lead to what it describes as a 'tidal wave' of nearly 277 million pieces of hate speech and other harmful content flooding onto Meta's platforms each year. [8]

5. Strong support for new conduct-based regulation

Our polling found strong support for additional regulatory measures that target the conduct and behaviour of senior leadership in social media firms. **Almost nine in 10 adults (87%) would support social media platforms being required to ensure managers with responsibility for user safety meet minimum standards of conduct.**

The recent actions of several tech firms have reignited debate about the need to build and strengthen the existing provisions set out in the Online Safety Act, including the value of a shift towards conduct-based regulation, as has been adopted in the financial services regime.

MRF sees particular value in requiring the staff and controlling owners of social media companies operating in the UK to be required to conduct themselves with due regard for the regulatory system and our rule of law. Financial services regulation has shown how regulation of conduct can achieve better outcomes than rules-based regimes.

There is a strong argument that the recent behaviour of tech platforms such as Meta actively demonstrates the need to regulate conduct, not simply to mandate a rules-based approach to the mitigation of individual and societal harms.

6 Meta's voluntary figures, issued in its Quarterly Transparency Reports, put these percentages even higher
7 Molly Rose Foundation (2024) How effectively do social networks moderate suicide and self-harm content? An analysis of the Digital Services Act Transparency Database
8 CCDH (2025) Meta Policy Changes Threaten 97% of its Hate Speech Enforcement

Next steps and recommendations

Our polling shows strong support for decisive action from both the government and regulator. When markets change fundamentally, how we regulate them should as well.

As it stands, the Online Safety Act is the best backstop we have to tackle the cavalier approach being adopted by Meta to public and user safety. However, there is nothing in the legislative existing framework to prevent Meta or indeed any other social media company from weakening their safety standards for adults, and Meta's changes highlight the limitations of Ofcom's current proposals to protect children.

What Government should do

The Government must act urgently to prevent a race to the bottom – with a clear recognition that Meta's rollback of its safety measures is only likely to be the start. Put simply, Meta's announcement marks a major strategic recalibration in its approach to protecting users, and this deeply retrograde step is a fundamental retrenchment from the objectives set out in the first clause of the Act.

The premise in which the original Act was passed has fundamentally changed. While the Secretary of State has said he will 'respond decisively' when new threats and risk types arise, his rhetoric is so far markedly out of step with his actions.

In light of Meta's changes, the Government must now act with urgency. We therefore call on the Secretary of State to:

Amend the OSA to include a 'no rollback' clause: the Government should urgently introduce a new duty on Category 1 services that prevents them from rolling back any policies or protections that were in place at the time that the Act was passed.[9]

Specify minimum standards of service: there is overwhelming support for social media platforms to provide a minimum standard of protection for users, and for UK speech rights to be determined by UK laws and regulators – rather than the commercial and political imperatives of foreign actors and governments.

Require proactive enforcement of policies relating to harmful content: strong proactive enforcement is a prerequisite to identify and remove harmful content at scale, and despite the claims from Meta, to maximise the safety and free expression rights of us all.

9 MRF was one of ten civil society signatories to a letter sent by the Online Safety Act Network to the Home Secretary and Technology Secretary that urged them to strengthen the Online Safety Act in response to Meta's changes. The letter can be read in full here: https://www.onlinesafetyact.net/analysis/meta-s-rollback-of-protections-for-users-why-the-uk-government-needs-to-act-and-fast/

What the regulator should do

We have been deeply concerned by Ofcom's reluctance to offer any public criticism of Meta's changes. Meta will almost certainly feel emboldened that Ofcom has chosen not to set out its public position, and it may even interpret this as a tacit greenlight for its safety rollbacks.

In late January, MRF wrote to Ofcom to encourage them to commit to a series of important changes, including bolstering the upcoming Protection of Children code and speeding up changes to its existing codes of practice.

There is little indication that the regulator will take any of these recommendations on board. As it stands, Ofcom's Protection of Children Code will leave children unnecessarily exposed to the impacts of Meta's safety downgrades – and this will send a powerful message to other companies about what they can get away with.

Ofcom must therefore act with urgency to:

- **State publicly that Meta's policy changes are inconsistent with the objectives of its Protection of Children Code:** By any reasonable measure, Meta's changes will impact the safety and free expression of multiple user groups, including women and girls, children with neurodiversity, young people experiencing poor mental health, and those identifying as LGBTQ+. Ofcom's silence on these impacts has been deafening.

- **Actively specify that Meta will be required to use proactive technology to enforce its obligations under the Protection of Children Code:** This must explicitly include the requirements to prevent children being algorithmically recommended content that might be harmful, including as a result of cumulative exposure to harmful material.

- **Commit to fast-track these and any other measures that are necessary to address the impact of Meta's policy and process changes:** Meta's announcement marks a major strategic recalibration in its approach to protecting users, and we cannot reasonably expect the regulator's current approach –a gradualist, iterative approach to strengthening its codes, with consultative cycles lasting up to 18 months – to be anything near adequate nor effective in response.

For a briefing and discussion about how we can work together to tackle preventable harm, please contact Andy Burrows: a.burrows@mollyrosefoundation.org

Registered Charity No: 1179482 https://mollyrosefoundation.org