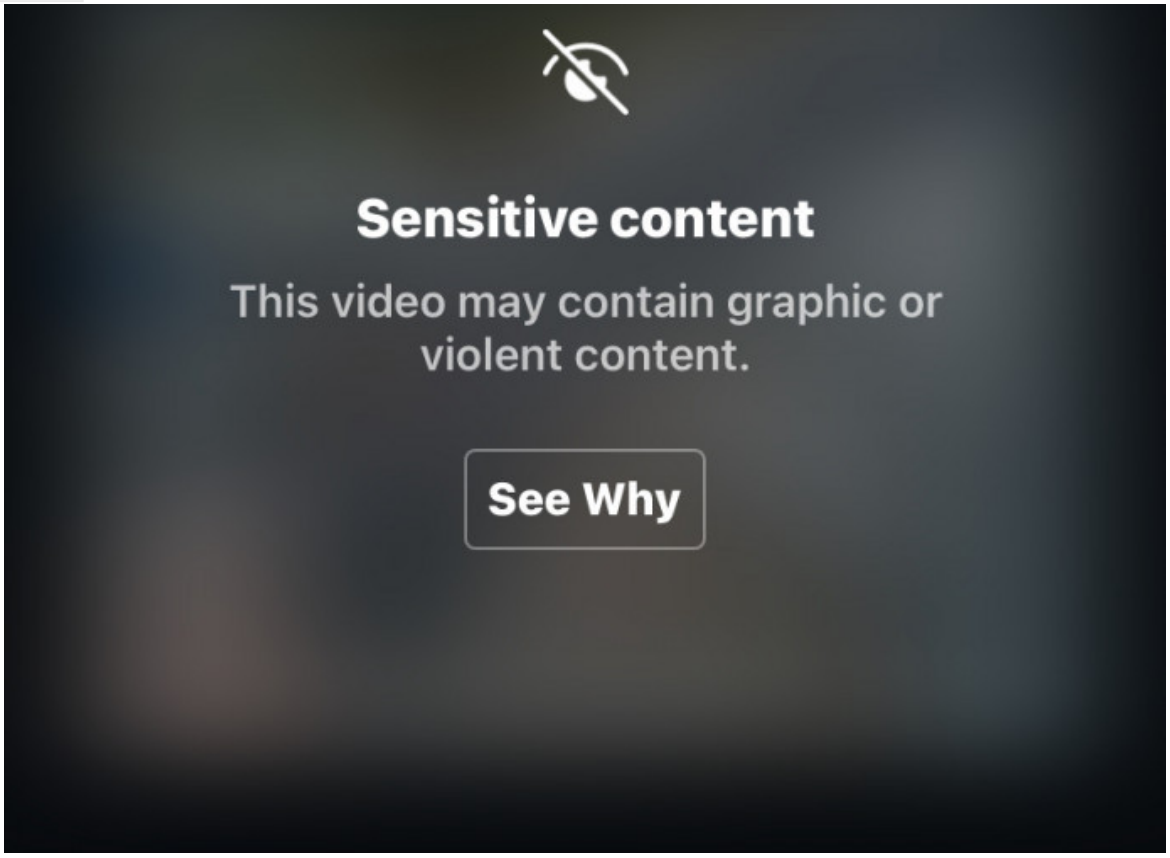
DAILY COMPASS
Managing Director Riccardo Cascioli
MADE FOR THE TRUTH

FACEBOOK

Meta aims for freedom of expression, but allows violence on the web

CULTURE 13_03_2025

Daniele Ciacci



Meta has been on a knife edge in recent days. It was only on 27 February that it fixed a technical glitch that caused violent content to appear in the personal Instagram Reels feeds of many users around the world. The company then apologised for the incident,

which occurred after a number of reports on social media of inappropriate content being displayed despite some users having activated the 'Sensitive Content Control' setting to filter out such material.

In a press release, Zuckerberg's company said it had managed to limit the spread of this violent content and apologised for any inconvenience caused by the issue. However, the Meta spokesperson was reluctant to divulge the most important information: what actually caused the problem. On this the company remained silent.

According to Meta's policy, particularly violent content should be removed at the source, including videos of dismemberment, visible internal organs and charred bodies. This filtering is supposed to be guaranteed by an artificial intelligence system capable of limiting it, albeit with some errors. In fact, some content will be allowed if it is useful for raising user awareness of issues such as human rights violations or armed conflict: in these cases, a warning will appear at the beginning of the reel, asking the user to confirm that he or she wishes to view content that could potentially upset them.

Significantly, however, this incident coincides with Meta's announcement that it is updating its moderation policy to promote greater freedom of expression. As of 7 January, the company has changed its automated systems to focus on 'illegal and serious violations' such as terrorism, child sexual exploitation and fraud, rather than 'all policy violations'. For less serious violations, Meta will rely more on user reports.

The most significant consequence of this new direction for Meta's moderation policy is the elimination of the fact-checking programme in the US on Facebook, Instagram and Threads, three of the largest social platforms with over 3 billion users worldwide.
Following this decision, Zuckerberg also stated that he had come under pressure from the Biden administration to censor content showing the side effects of Covid-19 vaccines.

Although this appears to be a turnaround aimed primarily at currying favour with Donald Trump, in recent years Meta has increasingly relied on its automated moderation tools to abandon the fact-checking that has often masked the inquisition of the powers that be.

So perhaps the mistake of recent days is nothing more than an unsuccessful attempt by Meta to effectively balance recommended content and user safety, without the constraint of fact-checkers' reports. This is why we are seeing the first balance

errors, such as the dissemination on Instagram of content on eating disorders aimed at teenagers.

It should also be noted that between 2022 and 2023, Meta reduced its workforce by around 21,000 employees, significantly reducing the groups responsible for information integrity and security. This unfortunate incident could therefore be the first stumbling block on a road that should free the tech company from the clutches of certain politics that is increasingly interested in exploiting it.