# Campaigners call on Ofcom to restrict AI use in Meta's social media risk assessments



Internet safety campaigners are increasingly vocal about their concerns regarding the use of artificial intelligence in risk assessments made by social media companies, particularly following reports of Meta's ambitious plans to automate up to 90% of these assessments. In a formal appeal to Ofcom, the UK’s communications regulator, groups including the Molly Rose Foundation, NSPCC, and Internet Watch Foundation have warned that AI-driven assessments could represent a “retrograde and highly alarming step” for online safety.

The Online Safety Act mandates social media platforms to thoroughly assess and mitigate potential harms, especially those affecting child users and illegal content. The process of risk assessment is thus a critical element in ensuring these platforms adhere to safety standards. Campaigners are apprehensive that reliance on AI could dilute the quality of these assessments, arguing that “risk assessments will not normally be considered as ‘suitable and sufficient’” if significantly automated.

As Ofcom weighs the implications of these warnings, it has stated that it is reviewing the concerns expressed and will provide a response in due course. The watchdog has consistently highlighted the importance of transparency, with a spokesperson confirming that companies should disclose who is responsible for conducting, reviewing, and approving their risk assessments.

In contrast, Meta has rejected the representations made by the campaigners, asserting that its risk management processes are mischaracterised. A spokesperson for the company asserted, “We are not using AI to make decisions about risk,” emphasising that their tools, developed by experts, aim to assist in identifying applicable legal and policy requirements rather than replacing human oversight. This assertion comes amidst claims from a former Meta executive, who indicated that the shift towards AI-driven assessments would expedite app updates but could potentially expose users to greater risks before problems are flagged.

Interestingly, Meta has also recently unveiled its Frontier AI Framework, which categorises AI systems based on their risk potential. In this framework, systems deemed too hazardous are to be indefinitely suspended, a step that reflects the company’s cautious approach to AI development. Yet, juxtaposed with this initiative, a study conducted by SaferAI has rated Meta's AI risk management efforts as “very weak”, raising questions about the robustness of its safety measures.

The evolving landscape of AI regulations is also a key factor in this debate. With the EU's AI Act coming into effect in August 2025, companies including Meta, Apple, and Microsoft are now actively enhancing their AI risk transparency to address emerging ethical, legal, and regulatory challenges. This legislation is poised to impose stricter regulations on AI, prompting a shift towards more transparent and responsible AI practices across the sector.

As these discussions unfold, the need for clear, effective oversight of AI's role in risk assessments remains crucial to safeguarding users, particularly the most vulnerable among them. The coming months are likely to be pivotal as Ofcom, Meta, and various advocacy groups navigate these complex concerns at the intersection of technology and safety.

### 📌 Reference Map:

* Paragraph 1 – [[1]](https://www.theguardian.com/technology/2025/jun/08/campainers-urge-uk-watchdog-to-limit-use-of-ai-after-report-of-meta-plan-to-automate-checks), [[2]](https://www.theguardian.com/technology/2025/jun/08/campainers-urge-uk-watchdog-to-limit-use-of-ai-after-report-of-meta-plan-to-automate-checks)
* Paragraph 2 – [[1]](https://www.theguardian.com/technology/2025/jun/08/campainers-urge-uk-watchdog-to-limit-use-of-ai-after-report-of-meta-plan-to-automate-checks), [[2]](https://www.theguardian.com/technology/2025/jun/08/campainers-urge-uk-watchdog-to-limit-use-of-ai-after-report-of-meta-plan-to-automate-checks), [[5]](https://www.thecable.ng/meta-to-suspend-ai-systems-with-severe-risks-under-new-guidelines/)
* Paragraph 3 – [[3]](https://about.fb.com/news/2025/02/meta-approach-frontier-ai/), [[6]](https://time.com/7026972/saferai-study-xai-meta/)
* Paragraph 4 – [[5]](https://www.thecable.ng/meta-to-suspend-ai-systems-with-severe-risks-under-new-guidelines/), [[7]](https://news.bloomberglaw.com/business-and-practice/meta-apple-microsoft-move-to-fend-off-mounting-ai-concerns)
* Paragraph 5 – [[4]](https://9to5mac.com/2025/02/04/meta-says-its-future-ai-models-could-have-catastrophic-outcomes/), [[6]](https://time.com/7026972/saferai-study-xai-meta/)

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.theguardian.com/technology/2025/jun/08/campainers-urge-uk-watchdog-to-limit-use-of-ai-after-report-of-meta-plan-to-automate-checks> - Please view link - unable to able to access data
2. <https://www.theguardian.com/technology/2025/jun/08/campainers-urge-uk-watchdog-to-limit-use-of-ai-after-report-of-meta-plan-to-automate-checks> - Internet safety campaigners have urged the UK's communications watchdog, Ofcom, to limit the use of artificial intelligence (AI) in crucial risk assessments. This follows a report that Meta, the owner of Facebook, Instagram, and WhatsApp, plans to automate up to 90% of its risk assessments using AI. Organisations such as the Molly Rose Foundation, the NSPCC, and the Internet Watch Foundation have expressed concerns, describing the move as a 'retrograde and highly alarming step'. They have called on Ofcom to publicly assert that risk assessments predominantly produced through automation will not be considered 'suitable and sufficient' under the UK's Online Safety Act. Ofcom has stated that it is 'considering the concerns' raised and will respond in due course.
3. <https://about.fb.com/news/2025/02/meta-approach-frontier-ai/> - Meta has introduced its Frontier AI Framework, outlining its approach to managing risks associated with advanced AI systems. The framework categorises AI systems into 'high risk' and 'critical risk' based on their potential to aid in cybersecurity, chemical, and biological attacks. Meta commits to halting the development of systems deemed too risky, aiming to prevent catastrophic outcomes. This initiative reflects Meta's cautious stance on AI development, emphasising safety and ethical considerations. ([about.fb.com](https://about.fb.com/news/2025/02/meta-approach-frontier-ai/?utm_source=openai))
4. <https://9to5mac.com/2025/02/04/meta-says-its-future-ai-models-could-have-catastrophic-outcomes/> - Meta has acknowledged the potential for its future AI models to lead to 'catastrophic outcomes'. In its Frontier AI Framework, Meta identifies 'high risk' and 'critical risk' systems, with the latter capable of causing large-scale, devastating, and potentially irreversible harm. The company outlines plans to prevent the release of such models, though it admits that complete mitigation may not always be possible. This admission underscores the challenges in AI risk management and Meta's commitment to safety. ([9to5mac.com](https://9to5mac.com/2025/02/04/meta-says-its-future-ai-models-could-have-catastrophic-outcomes/?utm_source=openai))
5. <https://www.thecable.ng/meta-to-suspend-ai-systems-with-severe-risks-under-new-guidelines/> - Meta has introduced a new artificial intelligence (AI) risk assessment framework that could lead to the suspension of AI systems deemed too dangerous. In its recent policy report titled, The Frontier AI Framework, Meta outlined a structured approach to evaluating potential risks. The company’s framework categorises AI risks into three levels: critical, high, and moderate. Meta said models posing the highest level of threat will not be developed or released. ([thecable.ng](https://www.thecable.ng/meta-to-suspend-ai-systems-with-severe-risks-under-new-guidelines/?utm_source=openai))
6. <https://time.com/7026972/saferai-study-xai-meta/> - A study by SaferAI has rated Meta's AI risk management as 'very weak'. The assessment evaluated companies based on their efforts to identify and mitigate AI risks, including technical evaluations and threat modelling. Meta, along with Mistral AI, received the lowest rating, indicating significant concerns about the company's approach to AI safety. This highlights the need for improved risk management practices in the rapidly evolving AI sector. ([time.com](https://time.com/7026972/saferai-study-xai-meta/?utm_source=openai))
7. <https://news.bloomberglaw.com/business-and-practice/meta-apple-microsoft-move-to-fend-off-mounting-ai-concerns> - In response to growing concerns over AI, companies like Meta, Apple, and Microsoft are enhancing their AI risk transparency. This includes setting up new roles such as chief AI ethics officer and establishing committees on AI ethics. The move aims to address ethical, legal, and regulatory issues associated with AI, including data privacy and the environmental impact of the technology. The EU's AI Act, effective August 1, 2025, will impose stricter regulations on AI systems, prompting companies to adopt more transparent and ethical AI practices. ([news.bloomberglaw.com](https://news.bloomberglaw.com/business-and-practice/meta-apple-microsoft-move-to-fend-off-mounting-ai-concerns?utm_source=openai))