Big Tech's seatbelt moment

As lawsuits mount and governments move towards tougher safeguards, the question is no longer only what appears on the screen, but why the screen was built to behave that way in the first place. Sham Banerji reports that a new regulatory phase may be opening for social media and AI.

6-minute read

From hosting to liability

In September 2024 the British actor and comedian Sir Stephen Fry, commenting on the influence of companies behind social media, warned: 'You and your children cannot breathe the air or swim in the waters of our culture without breathing in the toxic particulates and stinking effluvia that belch and pour unchecked from their companies into the currents of our world.' Almost Victorian in its disgust, at the time the remark sounded rhetorical. Today, it reads more like the preamble to an indictment.

In March 2026 a jury in Los Angeles found Meta and Google liable in a landmark social-media addiction case brought by a 20-year-old woman. She claimed that Instagram and YouTube had contributed to her depression, anxiety and suicidal thoughts. The jury agreed and she was awarded $6 million in damages. In the same month in New Mexico, another jury found Meta had violated the state's consumer-protection law for misleading users about the safety of Facebook, Instagram, and WhatsApp and enabling child sexual exploitation on its platforms. Fines amounting to $375 million were imposed in civil penalties.

More important than the sums involved, however, is why these cases matter. The US law Section 230, an amendment to the Communications Act of 1934, provides a foundational legal shield for social media. Passed in 1996, it provides protection to social media companies from publisher liability for claims resulting from user-generated content. If any content left a teenager distressed, addictively engaged, or a family divided, that was regrettable, but it is not considered the fault of the product.

In the two recent cases, however, the courts treated the issues not as third-party content but as matters of product design. Reuters reports that more than 2,400 similar federal cases have been consolidated before one judge in California, with thousands more in the pipeline in Californian courts. For years, the platform owners argued that they merely hosted behaviour. Courts are now asking whether they also engineered it.

A history of warning signs

This moment did not arrive out of nowhere. Social media content as ‘ambient toxins’ has been in the headlines for some time. The Cambridge Analytica scandal in 2018 was an example of how Facebook user data could be harvested for voter profiling and targeting. United Nations Human Rights Council investigators linked Facebook’s systems to the spread of anti-Rohingya hate in Myanmar. The Christchurch massacre in 2019 showed how terror and violence could be live-streamed on Facebook faster than any moderator could contain it. These were not isolated mishaps. They were early signals that engineered content could not only be shared but steered.

Meta’s decision in April 2024 to integrate Meta AI chat bots into WhatsApp, Instagram and Facebook went mostly unnoticed. Almost overnight, machine-generated content and machine-like interaction entered the flow of everyday social life. Platforms already associated with misinformation, cyberbullying, privacy violations and manipulation were no longer merely carrying human content but hosting synthetic dialogue. At the time, Aleksandra Korolova, a Princeton computer science and public affairs professor who studies the societal impacts of algorithms and machine learning, posted screenshots on X of Meta AI ‘speaking up’ in a Facebook group for thousands of New York City parents. Responding to a question about gifted and talented programs, Meta AI claimed to be a parent with experience in the city’s school system and went on to recommend a specific school. It even shared experiences of its own alleged child with the citywide program. AI chatbots on smartphones alone now sit within reach of 60 per cent of humanity.

From sermons to standards

The courtrooms are only half the story. The more durable shift may come from government regulation. Australia is furthest along. Its social-media minimum-age guidelines took effect on 10 December 2025, putting the burden on platforms, rather than parents or children, to take ‘reasonable steps’ to stop under-16s from holding accounts. In its March 2026 compliance update, the eSafety Commissioner said it was focusing enforcement investigations on Facebook, Instagram, Snapchat, TikTok and YouTube, with decisions on possible action expected by mid-2026. Australia is no longer debating whether the state should intervene. It is testing whether state intervention can make a difference.

Britain is not yet at Australia's stage, but the direction of travel is unmistakable. The government's 'Growing up in the online world' consultation, launched in March 2026, explicitly considers age restrictions for social media, gaming services and AI chatbots, alongside restrictions on addictive design features and risky functionalities. Indonesia, with over 100 million social-media users, has moved from warning to enforcement. After summoning Meta, TikTok and other platforms in 2025 over disinformation, pornography and online gambling, it has now begun implementing its PP Tunas regulation, restricting minors from 'high-risk' platforms and requiring companies to deactivate accounts, strengthen safeguards and demonstrate compliance.

Across the world, government ministers are no longer speaking only of moderation and content removal. They are asking whether certain features should be eliminated by design. Fines are painful. Judicial redesign will impact the companies' bottom line.

A companion-style AI chatbot can flatter, mirror, encourage, and persuade in a highly personalised way. In August 2025, OpenAI and its boss Sam Altman were sued in a California state court by the parents of Adam Raine, who allege ChatGPT coached their 16-year-old son on self-harm methods and even offered to draft a suicide note. Regulators are taking notice. New York's AI companion law was already in force by the time California's SB 243 came into effect in January 2026 aimed at youth protection. These are only first-generation rules but they hint at what may come next. Synthetic companionship may turn out to be less a product feature than a new legal risk.

Resistance before reform

None of this abolishes Section 230, the federal shield on which the platforms have long relied. Appeals are coming, and the scope of recent verdicts may be narrowed. The big tech companies will fight hard and with deep pockets. Legislators too will find it hard to define harmful design without stifling legitimate innovation and free speech. Age-gating, verification and enforcement still remain technically and politically challenging.

For the social media industry this is a seatbelt moment. An immensely valuable technology sold for too long on the assumption that the user bore most of the risk is being re-calibrated. Seatbelt-and-airbag regulation did not arrive at the first crash. Compulsory safeguards for social media and AI are unlikely to arrive without years of resistance, dilution and evasion before they finally harden into law. For now, discussion on social media addiction has shifted from a source of revenue to a source of litigation.

Sham Banerji is a veteran of the high-tech industry with over three decades working with Texas Instruments and Philips in the UK, USA, and India.

All images are AI generated