Meta CTO thinks bad metaverse moderation could pose an ‘existential threat’

Meta (formerly Facebook) CTO Andrew Bosworth warned employees that creating safe virtual reality experiences was a vital part of its business plan — but also potentially impossible at a large scale.

In an internal memo seen by *Financial Times*, Bosworth apparently said he wanted Meta virtual worlds to have “almost Disney levels of safety,” although spaces from third-party developers could have looser standards than directly Meta-built content. Harassment or other toxic behavior could pose an “existential threat” to the company’s plans for an embodied future internet if it turned mainstream consumers off VR.

At the same time, Bosworth said policing user behavior “at any meaningful scale is practically impossible.” *FT* reporter Hannah Murphy later tweeted that Bosworth was citing Masnick’s Impossibility Theorem: a maxim, coined by *Techdirt* founder Mike Masnick, that says “content moderation at scale is impossible to do well.” (Masnick’s writing notes that this isn’t an argument against pushing for better moderation, but large systems will “always end up frustrating very large segments of the population.”)

Virtual worlds could have “a stronger bias towards enforcement”

Bosworth apparently suggests that Meta could moderate spaces like its Horizon Worlds VR platform using a stricter version of its existing community rules, saying VR or metaverse moderation could have “a stronger bias towards enforcement

along some sort of spectrum of warning, successively longer suspensions, and ultimately expulsion from multi-user spaces."

While the full memo isn't publicly available, Bosworth posted a blog entry alluding to it later in the day. The post, titled "Keeping people safe in VR and beyond," references several of Meta's existing VR moderation tools. That includes letting people block other users in VR, as well as an extensive Horizon surveillance system for monitoring and reporting bad behavior. Meta has also pledged $50 million for research into practical and ethical issues around its metaverse plans.

As *FT* notes, Meta's older platforms like Facebook and Instagram have been castigated for serious moderation failures, including slow and inadequate responses to content that promoted hate and incited violence. The company's recent rebranding offers a potential fresh start, but as the memo notes, VR and virtual worlds will likely face an entirely new set of problems on top of existing issues.

"We often have frank conversations internally and externally about the challenges we face, the trade-offs involved, and the potential outcomes of our work," Bosworth wrote in the blog post. "There are tough societal and technical problems at play, and we grapple with them daily."

https://www.theverge.com/2021/11/12/22779006/meta-facebook-cto-andrew-bosworth-memo-metaverse-disney-safety-content-moderation-scale