Satya Nadella says we should assume all AI models are ‘compromised’

In a lengthy post on X, Microsoft’s CEO laid out his views on the dangers posed by highly advanced AI models and how to confront those risks. Nadella says we can no longer accept a world where AI is treated as a “set of nested black boxes” whose advice and actions we simply accept or reject. He calls for building a more transparent system where models can be contained, observed, and leaves behind “tamper-proof human readable evidence.”

Many of his recommendations align with what we’ve heard from others in the industry: timely incident disclosure, independent audits, verifiable data, and containment. It’s on that last point that he appears t …

Read the full story at The Verge.

https://www.theverge.com/ai-artificial-intelligence/1009337/satya-nadella-says-we-should-assume-all-ai-models-are-compromised