Detecting Online Manipulation

1 min

The combination of generation algorithms and the speed at which information moves, facilitated by various social media services, news, and other outlets, has led to an explosion of misinformation. While not all directly attributable to AI, the ability to create large amounts of content pushing a similar narrative or create entirely high-quality false content significantly threatens the overall state of information.

Given this, detecting this content is now more critical than ever.

When it comes to detecting fake content, we must ensure our information is sourced from verified and reputable sources. New/disreputable sources may be more apt to spread and share false AI-created content. Identifying this early on is essential to ensure we have a truthful narrative.

Given this issue's importance, many organizations have also implemented policies to further aid users in identifying AI-related content. For example, Google now requires advertisers to disclose when AI is used in their content. It is essential to watch for this as it can further help identify when AI may be used for misleading purposes.

Research is also being focused on using AI to identify AI-generated content. Unfortunately, as generative models continue to train themselves, many AI models struggle to keep up. For example, OpenAI, the developers behind ChatGPT, removed their detection tool due to low performance. While some tools remain available, they should be approached with caution as they may not perform as expected.

As always, the easiest way to identify this content is to cross-reference data and double-check the content you may be reviewing.