observer.co.uk

‘AI swarms’ are mass-producing credible misinformation. Democracy may get stung

John Naughton

5–7 minutes

How much of what goes on in social media is real – that is, consisting of genuine conversations between humans rather than interactions with bots? The answer is that nobody really knows, which is worrying because the erosion of trust between what’s real and what’s manufactured is making democracies ungovernable.

Take X (nee Twitter), which is still the most powerful platform around. If you ask what percentage of X accounts are fake or robotic, the service won’t tell you because it doesn’t publicly disclose precise bot versus real account breakdowns. So we have to fall back on estimates from independent studies and analyses. These vary widely as a result of differing methodologies and definitions of “bots”. A few have put it that bots make up about 5% of users but other estimates vary so wildly – from 20% to 64% – that they reveal how little we actually know, which is itself part of the problem.

The reticence of X about internal data is par for the industry course, by the way. Tech companies are pathologically shy about their internal data – remember what Meta knew about harms to young users but chose not to disclose.

To be fair, there are genuine terminological puzzles here because the phrase “bots on X” lumps very different kinds of accounts together. First of all, there are old-style, fully automated bots that post on a fixed schedule and automatically reply, retweet, or like. Then there are “cyborgs”, where a human runs the account but automation handles posting, liking, following or replies. These are often coordinated with many similar accounts and used for political or marketing campaigns or for boosting individuals.

They were the tools used, for example, to boost the rightwing American extremist Nick Fuentes on X to the level where mainstream media began to imagine that he must be a serious figure. They were also used by a propaganda operation run from Russia that included nearly 1,000 accounts pretending to be Americans who were posting pro-Russia stories on the platform.

The next escalation in this process of manufacturing “reality” is now upon us, courtesy of AI. A recently published paper by a large group of scholars in the prestigious journal *Science* lays out the scenario. ChatGPT et al offer the prospect of manipulating beliefs and behaviours on “a population-wide level”. The combination of large language models (LLMs) and autonomous agents will enable

what the researchers call “AI swarms” to reach “unprecedented scale and precision”. They will expand propaganda output without sacrificing credibility and inexpensively create “falsehoods that are rated as more humanlike than those written by humans”.

These capabilities easily transcend the limitations of the “dumb” botnets favoured by the Russians, Chinese and others, which simply amplified the spread of misinformation by incessantly retweeting to trigger algorithmic visibility through repetition, manual scheduling and rigid scripts.

An AI swarm is fundamentally different: it maintains persistent identities and memory, coordinates towards shared objectives while varying tone and content and, crucially, “adapts in real time to engagement, platform cues, and human responses; operates with minimal human oversight; and can deploy across platforms”.

Related articles:

At the core of this capability is its ability to exploit the Eliza effect; the tendency of people to attribute humanlike understanding and emotions to machines that appear to have conversational skills – and to trust those machines as a result. We’re now seeing this all over the place with LLMs; for example, in the volume of people turning to ChatGPT for health advice, or the vast numbers apparently using LLMs for help with emotional or mental health problems. Given this, the subtler propaganda enabled by LLMs may have a greater effect on people than the cruder assertions and bland claims characteristic of old-style online campaigns.

Those campaigns were based on harnessing social media recommendation algorithms to target messages at *individuals*. The wider significance of this new use of AI is its potential to operate on a much broader scale. Smart bot swarms, the researchers argue, may be capable of engineering a synthetic consensus by seeding narratives across disparate echo chambers that normally don’t communicate, creating an illusion of majority agreement that spans political divides.

And they could also boost this illusion by liking posts, making narratives appear widely supported and creating “a mirage of bipartisan grassroot consensus with enhanced speed and persuasiveness. The result is deeply embedded manipulation that lets operators nudge public discourse almost invisibly over time.”

We may thus be looking at a new way of doing what mass media did in earlier ages – what Noam Chomsky (and, before him, Walter Lippmann) described as “manufacturing the consent” of the governed to whatever their political masters have decided to

do. Dystopia 2.0, here we come.

What I’m reading

Ulysses AI lecture

Top Trumps

Hands-off approach

Photograph by Lee Lockwood/Getty Images