Die nächste Generation an Desinformation

Bald schon könnten neue "KI-Schwärme" einen öffentlichen Konsens vortäuschen und die Demokratie im Stillen verzerren, warnen Forschende. David Garcia von der Universität Konstanz gibt einen Ausblick auf ein noch nie dagewesenes Ausmaß an Meinungsmanipulation, das den demokratischen Diskurs bedroht.

Die nächste Generation an politischer Einflussnahme im Internet werde vermutlich nicht länger aus offensichtlichen "Copy-and-Paste-Bots" bestehen, sondern aus koordinierten Gemeinschaften von "KI-Schwärmen": Hiervor warnt ein internationales Forschungsteam unter Beteiligung des Konstanzer Social Data-Forschers David Garcia. Gemeint sind Flotten von KI-gesteuerten Personas, die sich in Echtzeit anpassen, Gruppen infiltrieren und in großem Umfang einen Anschein von gemeinschaftlicher Meinung erwecken können. Ein Chor aus scheinbar unabhängigen Stimmen schafft die Illusion eines breiten öffentlichen Konsenses, während er in Wirklichkeit Desinformation verbreitet. In einem Artikel in der renommierten Fachzeitschrift Science beschreiben die Autor*innen, wie es durch die Verschmelzung von großen Sprachmodellen (LLMs) mit Multiagentensystemen zu "schädlichen KI-Schwärmen" kommen könnte, die soziale Dynamiken authentisch imitieren - und den demokratischen Diskurs bedrohen, indem sie falsche Tatsachen zementieren und Konsens suggerieren.

Gefälschter Konsens

Das Forschungsteam zeigt auf, dass die zentrale Gefahr nicht nur in falschen Inhalten besteht, sondern vor allem in einem "künstlichen Konsens": Der falsche Eindruck, dass "ja jeder das sage", kann Überzeugungen und Normen beeinflussen, selbst wenn einzelne Behauptungen umstritten sind. Dieser anhaltende Einfluss, so die Forschenden, könne tiefgreifende kulturelle Veränderungen bewirken, die über Normenverschiebungen hinausgehen und die Sprache, Symbole und Identität einer Gemeinschaft auf subtile Weise verändern.

"Die Gefahr besteht nicht mehr nur in Fake News, sondern darin, dass die Grundlage des demokratischen Diskurses - unabhängige Stimmen - zusammenbricht, wenn ein einzelner Akteur Tausende von einzigartigen, KI-generierten Profilen kontrollieren kann", schildert Jonas R. Kunst (BI Norwegian Business School), einer der Hauptautoren des Artikels.

Darüber hinaus können diese schädlichen KI-Schwärme auch die Trainingsdaten von regulärer künstlicher Intelligenz verunreinigen, indem sie das Internet mit gefälschten Behauptungen überfluten. Auf diese Weise können sie ihren Einfluss auf etablierte KI-Plattformen ausweiten. Die Forscher*innen warnen, dass diese Bedrohung nicht nur theoretisch ist: Analysen legen nahe, dass derartige Taktiken bereits angewendet werden.

Was ist ein KI-Schwarm?

Die Forscher*innen definieren einen schädlichen KI-Schwarm als eine Gruppe von KI-gesteuerten Akteuren, die dauerhafte Identitäten bewahren und ein Gedächtnis haben, sich auf gemeinsame Ziele koordinieren und dabei Ton und Inhalt variieren. Sie passen sich in Echtzeit an Interaktionen und menschliche Reaktionen an, benötigen nur minimale Aufsicht durch Menschen und können plattformübergreifend eingesetzt werden. Im Vergleich zu früheren Bot-Netzen könnten solche Schwärme schwieriger zu erkennen sein, da sie heterogene, kontextbezogene Inhalte generieren und sich dennoch in koordinierten Mustern bewegen.

Strategien gegen KI-Schwärme

"Über die Täuschungen oder die Sicherheit von einzelnen Chatbots hinaus müssen wir neue Gefahren erforschen, die sich aus der Interaktion von vielen KI-Akteuren ergeben. Dazu ist es essenziell, dass wir diese KI-Akteure mit Methoden der Verhaltenswissenschaften untersuchen und ihr kollektives Verhalten analysieren, wenn sie in großen Gruppen interagieren?, betont David Garcia, Professor für Social and Behavioural Data Science an der Universität Konstanz.

Statt einzelne Beiträge zu moderieren, plädieren die Forscher*innen für Schutzmaßnahmen, die koordiniertes Verhalten und die Herkunft der Inhalte verfolgen: statistisch unwahrscheinliche Muster von Koordinierung aufdecken, Verifizierungsoptionen unter Wahrung des Datenschutzes anbieten und Hinweise auf KI-Einflussnahme über verteilte Beobachtungszentren weitergeben. Gleichzeitig sollten Anreize verringert werden, indem die Monetarisierung von gefälschten Interaktionen eingeschränkt und die Rechenschaftspflicht erhöht werden.

Originalpublikation:

Originalpublikation: Daniel Thilo Schroeder et al., How malicious AI swarms can threaten democracy. Science 391, 354-357 (2026).
DOI: 10.1126/science.adz1697

Link: https://www.science.org/doi/10.1126/science.adz1697