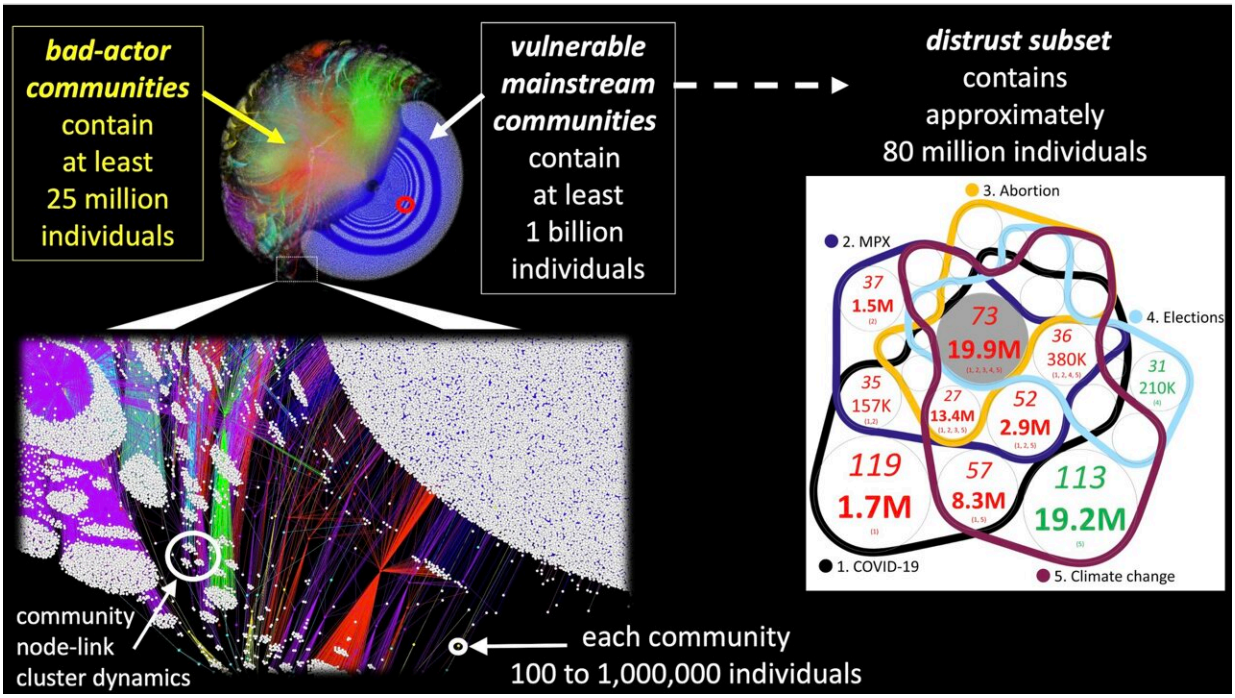
New research addresses predicting and controlling bad actor AI activity in a year of global elections

January 23 2024



The bad-actor–vulnerable-mainstream ecosystem (left panel). It comprises interlinked bad-actor communities (colored nodes) and vulnerable mainstream communities (white nodes, which are communities to which bad-actor communities have formed a direct link). This empirical network is shown using the ForceAtlas2 layout algorithm, which is spontaneous, hence sets of communities (nodes) appear closer together when they share more links. Different colors correspond to different platforms. Small red ring shows 2023 Texas shooter's YouTube community as illustration. Right panel shows Venn diagram of the topics discussed within the distrust subset. Each circle denotes a

category of communities that discuss a specific set of topics, listed at bottom. The medium size number is the number of communities discussing that specific set of topics, and the largest number is the corresponding number of individuals, e.g. gray circle shows that 19.9M individuals (73 communities) discuss all 5 topics. Number is red if a majority are anti-vaccination; green if majority is neutral on vaccines. Only regions with > 3% of total communities are labeled. Anti-vaccination dominates. Overall, this figure shows how bad-actor-AI could quickly achieve global reach and could also grow rapidly by drawing in communities with existing distrust. Credit: Johnson et al.

More than 50 countries are set to hold national elections this year and analysts have long sounded the alarm on the threat of bad actors using artificial intelligence (AI) to disseminate and amplify disinformation during the election season across the globe.

Now, a new study led by researchers at the George Washington University predicts that daily bad-actor AI activity will escalate by mid-2024, increasing the threat that it could affect election results. The research is the first quantitative scientific analysis that predicts how bad actors will misuse AI globally.

The paper, "Controlling bad-actor-AI activity at scale across online battlefields," is published in the journal *PNAS Nexus*.

"Everybody is talking about the dangers of AI, but until our study there was no science of this threat," Neil Johnson, lead study author and a professor of physics at GW, says. "You cannot win a battle without a deep understanding of the battlefield."

The researchers say the study answers the what, where, and when AI will be used by bad actors globally, and how it can be controlled. Among their findings:

- Bad actors need only basic Generative Pre-trained Transformer (GPT) AI systems to manipulate and bias information on platforms, rather than more advanced systems such as GPT 3 and 4, which tend to have more guardrails to mitigate bad activity.
- A road network across 23 social media platforms, which was previously mapped out in Johnson's prior research, will allow bad actor communities direct links to billions of users worldwide without users' knowledge.
- Bad-actor activity driven by AI will become a daily occurrence by the summer of 2024. To determine this, the researchers used proxy data from two historical, technologically similar incidents that involved the manipulation of online electronic information systems. The first set of data came from automated algorithm attacks on U.S. financial markets in 2008, and the second came from Chinese cyber attacks on U.S. infrastructure in 2013. By analyzing these data sets, the researchers were able to extrapolate the frequency of attacks in these chains of events and examine this information in the context of the current technological progress of AI.
- Social media companies should deploy tactics to contain the disinformation, as opposed to removing every piece of content. According to the researchers, this looks like removing the bigger pockets of coordinated activity while putting up with the smaller, isolated actors.

More information: Neil F Johnson et al, Controlling bad-actor-artificial intelligence activity at scale across online battlefields, *PNAS Nexus* (2024). DOI: 10.1093/pnasnexus/pgae004.
academic.oup.com/pnasnexus/art … /7582771?login=false

Provided by George Washington University