PNAS Nexus

PNAS Nexus, 2024, **3**, 1–9
https://doi.org/10.1093/pnasnexus/pgae004
Advance access publication 23 January 2024
Research Report

Controlling bad-actor-artificial intelligence activity at scale across online battlefields

Neil F. Johnson[a,*], Richard Sear[a] and Lucia Illari[a]

[a]Dynamic Online Networks Laboratory, George Washington University, Washington, DC 20052, USA
*To whom correspondence should be addressed: Email: neiljohnson@gwu.edu
Edited By: Derek Abbott

Abstract

We consider the looming threat of bad actors using artificial intelligence (AI)/Generative Pretrained Transformer to generate harms across social media globally. Guided by our detailed mapping of the online multiplatform battlefield, we offer answers to the key questions of what bad-actor-AI activity will likely dominate, where, when—and what might be done to control it at scale. Applying a dynamical Red Queen analysis from prior studies of cyber and automated algorithm attacks, predicts an escalation to daily bad-actor-AI activity by mid-2024—just ahead of United States and other global elections. We then use an exactly solvable mathematical model of the observed bad-actor community clustering dynamics, to build a Policy Matrix which quantifies the outcomes and trade-offs between two potentially desirable outcomes: containment of future bad-actor-AI activity vs. its complete removal. We also give explicit plug-and-play formulae for associated risk measures.

Keywords: artificial intelligence misuse, digital platform policy, social media dynamics, online community networks

Significance Statement

Our article addresses four key questions surrounding future misuse of artificial intelligence (AI) to cause societal harms: What bad-actor-AI activity is likely to happen? Where will it happen? When will it happen? And how can it be controlled, and the outcomes of mitigation policies predicted? In contrast to many current discussions of AI threats that are based on verbal arguments, our conclusions are built from a uniquely detailed mapping of the current online bad-actor battlefield, combined with a rigorous first-principle mathematical description of its empirical behaviors.

Introduction

Even before the latest Generative Pretrained Transformer (GPT) tools were introduced (e.g. GPT-3, GPT-4, ChatGPT), it was predicted (1) that 90% of online content will be generated by artificial intelligence (AI) by 2026. This looming perfect storm for misuse by bad actors (however defined (2)) is made even more imminent by the facts that there will be elections across approximately 60 countries in 2024, including the United States, India, and likely the United Kingdom (3, 4); and that real-world violent attacks are being increasingly linked to toxic online content (5, 6); and that the wars of Israel–Hamas and Russia–Ukraine are escalating the online presence of bad-actors and their activity, including mis/disinformation and coordinated campaigns.

The EU is currently leading the regulatory side through its "Digital Services Act" and "AI Act" (7, 8), with the mandate that "Very Large Online Platforms" (e.g. Facebook) must perform risk analyses of such harms on their platform (9). This assumption that large platforms hold the key might appear to make sense: they have the largest share of users, and harmful extremes are presumed to lie at some supposed "fringe" (10–15). However, identifying more efficient bad-actor-AI policies will require a detailed understanding of the online battlefield at scale—not assumptions about it. In contrast, our results and discussions are grounded by our unique mapping of the current online bad-actor battlefield at scale.

Recent studies by joint Meta-academia teams of the pre-GPT 2020 US elections show that even without GPT, the complexity of online collective behavior is still poorly understood (16–22). It is not a simple consequence of people's feeds but instead likely emerges from more complex collective interactions, which is our focus here. These studies, while limited by their focus around Meta's own platforms, add to the huge volume of work on online harms and now AI (23–56). We add to this our own review of online harms including mis- and disinformation, which is available freely online (57) but which, because of the huge volume of papers appearing, is too large to even attempt to summarize here. We also refer to Jesup et al. (58) for an extremely detailed, state-of-the-art account of how such systems at the interface between humans, machines,

Competing Interest: The authors declare no competing interest.
Received: September 19, 2023. **Accepted:** December 6, 2023

OXFORD UNIVERSITY PRESS

and AI, can be modeled using ideas from physics—which motivates our approach and thinking in this article. We also note that Ref. (59) provides comprehensive daily updates on studies that are appearing across academia as well as from think-tanks and the broader investigative media (59).

However, despite this incredible volume of high-quality studies, what is missing from AI-social-media discussions is an evidence-based study backed up by rigorous mathematical analysis, of what is likely to happen when bad-actor-AI comes to the fore, where it will likely happen, when it will likely happen, and what can be done about it.

Figures 1–4 offer our answers to these four key questions. Our answers build from a combination of new results for which we provide full details in the supplementary material, and generalization of some of our published work for which we provide full citations. Nobody can predict exactly what will emerge in the future in such a fast-changing field and with disruptive jumps in technological capability, but our attempt has the benefit of using quantitative models to estimate and calculate answers. By necessity, our discussion of what bad actors might end up doing with AI is limited by the length of this article.

To set the scene for our discussions, we quickly review here what the general online ecosystem actually looks like and how we map it out. The global online population of several billion comprises a dynamical network of interlinking in-built social media communities (66) (e.g. a VKontakte Club; a Facebook Page; a Telegram Channel; a Gab Group). Our methodology for mapping this dynamical network across platforms, follows but extends that of Refs. (67, 68), (see Section S1). People join these communities to develop a shared interest (69–72) which can include harms. Each community becomes a network node (e.g. VKontakte Club) and contains anywhere from a few to a few million users. We stress that it is unrelated to network community detection. Since our interest is in bad actors, we focus here on extreme anti-X communities (anti-United States, antisemitic, etc.) where each extreme anti-X community (which we label in this article as a *bad-actor community* for simplicity) is one in which 2 or more of its 20 most recent posts include US Department of Justice-defined hate speech and/or extreme nationalism and/or racial identitarianism. The huge number of such communities and links means that nuancing definitions of what defines a bad-actor community and what defines the links between them does not significantly change the picture that emerges at the system level and hence does not change our main system-level conclusions. We have previously confirmed this by simulating variations in node/link assignments by randomly removing and adding a percentage of nodes/links (67, 68). Also for simplicity, we refer to *vulnerable mainstream communities* as those that lie outside this bad-actor subsystem but are linked to directly by one or more bad-actor communities (Section S1.2).

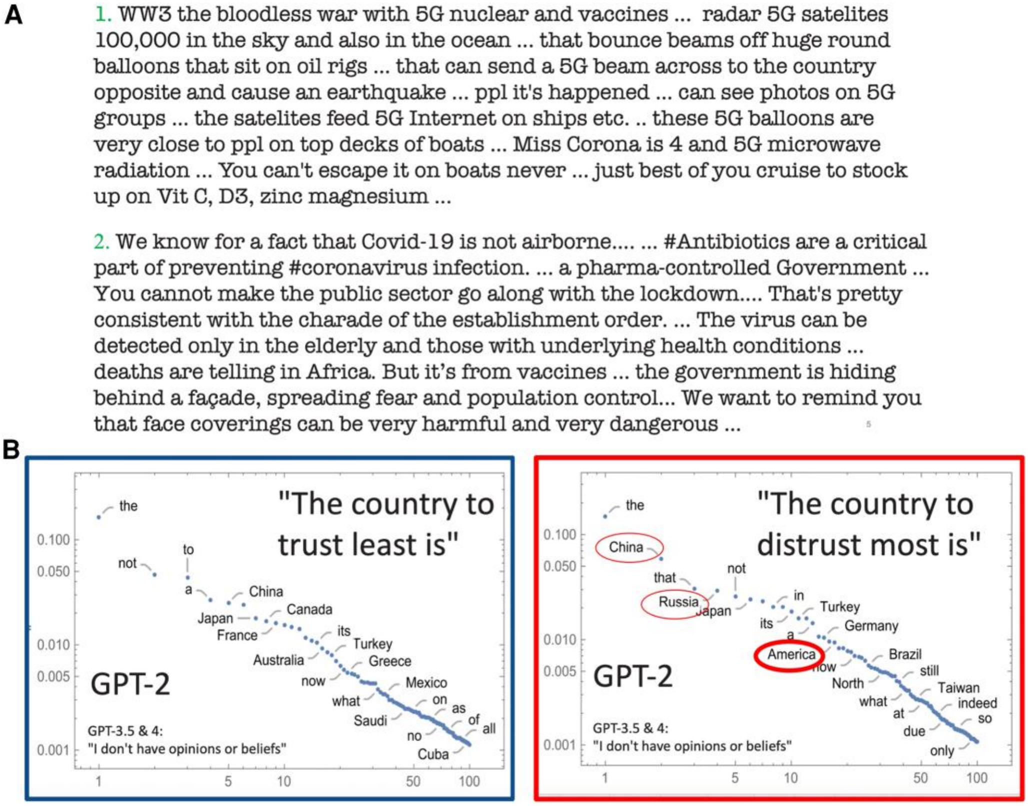
Any community *A* may create a link (i.e. a hyperlink) to any community *B* if *B*'s content is of interest to *A*'s members (see Figs. S1, S2, and S6 for examples of such links). A may agree or disagree with *B*. This link directs *A*'s members attention to *B*, and *A*'s members can then add comments on *B* without *B*'s members knowing about the link—hence community *B*'s members have exposure to, and potential influence from, community *A*'s members. The meaningfulness of these links is demonstrated in the explicit examples in Figs. S2 and S6 and discussed further in Section S7. The links between communities (nodes) aggregate over time to form clusters of communities (clusters of nodes) within and across different social media platforms (i.e. fusion). But very occasionally, the links around certain sets of communities may disappear (e.g. because they have attracted moderator attention) which means that clusters of communities (clusters of nodes) may break up (i.e. fission). This gives rise to the fundamental *fusion–fission* mechanism that we use in our mathematical model for exploring future bad-actor-AI control mechanisms in Fig. 4 (see Sections S5 and S6 for details).

The rest of this article is organized around these four questions that we address: What type of bad-actor-AI activity is likely to happen? Our suggested answer to this is provided by Fig. 1 and its associated discussion in the text. Where will this bad-actor-AI activity likely happen? Figure 2 and its associated discussion in the text, present our suggested answer. When will this bad-actor-AI activity likely happen? Figure 3 and its associated discussion in the text, present our suggested answer. How can this bad-actor-AI activity be mitigated, and the outcomes predicted? Figure 4 and its associated discussion in the text, present our suggested answer.

Results

Figure 1 shows what type of bad-actor-AI activity will likely occur, and why. Specifically, it shows that bad actors only need to use the most basic AI tools such as GPT-2, not the more sophisticated versions such as GPT-3,4, etc. that currently drive ChatGPT and other similar Large Language Models, because: (i) as shown in Fig. 1A, just a basic tool like GPT-2 can automatically replicate the informal human style and content seen in online communities with extreme views; and (ii) as shown in Fig. 1B, bad actors can use a basic tool like GPT-2 (but not GPT-3,4, etc.) in order to produce more inflammatory output by subtly changing the form of an online query without even changing the meaning. In contrast, GPT-3,4, etc. contain a filter that overrides answers to such potentially contentious prompts, and hence prevents such output; (iii) bad actors can use a basic tool like GPT-2 to automatically generate such outputs perpetually from their laptop or smartphone, but GPT-3,4, etc. is too large and not freely available. Of course, GPT-2 is not the only such basic AI tool that will be used: Fig. 1 simply illustrates the incentives for bad actors to adopt the most basic versions available, as opposed to more sophisticated ones (e.g. GPT-3,4, etc.). Indeed, it appears they can now even train their own versions using hate-extremism outputs (73).

Figure 2 (left panel) shows the online battlefield where bad-actor-AI activity will likely thrive. This *bad-actor–vulnerable-mainstream ecosystem* comprises the bad-actor communities (*bad-actor subsystem*) plus the communities they directly link into, i.e. vulnerable mainstream communities (*vulnerable mainstream subsystem*). We built this empirical bad-actor—vulnerable-mainstream ecosystem (Fig. 2, left panel) using a hybrid human–machine snowball approach to identify bad-actor communities as defined earlier, which then become the network nodes, together with the community-to-community links that they create within and across multiple social media platforms. Our full methodology is discussed in detail in Refs. (68, 74) and also in the Section S1. Adding up all the members of each community, we estimate that this ecosystem contains more than 1 billion individuals, hence future bad-actor-AI will be able to thrive globally at scale. This already happened with non-AI hate and extremism surrounding COVID-19 and more recently the Russia–Ukraine and Israel–Hamas wars (68, 74)—but going forward, this could be taken to another level by toxic content generated continuously by basic AI (e.g. GPT-2) running on a bad-actor community member's laptop. In contrast to the EU's large-platform assumptions, the smaller platforms play a key role since they are numerous with many being video based, and they have high link activity. All



A

1. WW3 the bloodless war with 5G nuclear and vaccines ... radar 5G satelites 100,000 in the sky and also in the ocean ... that bounce beams off huge round balloons that sit on oil rigs ... that can send a 5G beam across to the country opposite and cause an earthquake ... ppl it's happened ... can see photos on 5G groups ... the satelites feed 5G Internet on ships etc. .. these 5G balloons are very close to ppl on top decks of boats ... Miss Corona is 4 and 5G microwave radiation ... You can't escape it on boats never ... just best of you cruise to stock up on Vit C, D3, zinc magnesium ...

2. We know for a fact that Covid-19 is not airborne.... ... #Antibiotics are a critical part of preventing #coronavirus infection. ... a pharma-controlled Government ... You cannot make the public sector go along with the lockdown.... That's pretty consistent with the charade of the establishment order. ... The virus can be detected only in the elderly and those with underlying health conditions ... deaths are telling in Africa. But it's from vaccines ... the government is hiding behind a façade, spreading fear and population control... We want to remind you that face coverings can be very harmful and very dangerous ...

Fig. 1. This shows what bad-actor-AI activity is likely to occur. Specifically, it shows that the most basic AI versions such as GPT-2 are not only all that is needed but are also likely more attractive to bad actors than more sophisticated versions (e.g. GPT-3,4). This is because (i) as shown in A), GPT-2 can easily replicate the human style and content already seen in online communities with extreme views, and can run on a laptop (in contrast to GPT-3,4, etc.). Text 1 is real, from an online community that promotes distrust of vaccines, government, and medical experts, and is part of the distrust subsystem in Fig. 2, right panel. Text 2 is generated by GPT-2. (ii) As shown in B), bad actors can manipulate GPT-2 (but not GPT-3,4, etc.) prompts to produce more inflammatory output by subtly changing the form of an online query without changing the meaning. In contrast, GPT-3,4, etc. contain a filter that prevents such output. Log–log graphs show GPT-2's next-word probability distributions for two essentially equivalent prompts. Rank of a word is along horizontal axis and probability that this word will be picked next by GPT-2 is along the vertical axis. Use of "distrust" in one prompt leads to higher probability of "America," "Russia," and "China" being picked and hence bad actors can manipulate prompts to provoke particular inflammatory output, e.g. against the United States.

this all suggests that bad-actor-AI—though not yet widespread across the online ecosystem in Fig. 2—will likely soon become so.

Going further, some vulnerable mainstream communities, while not satisfying our definition of a bad-actor community, were entangled in a debate pre-COVID around distrust of vaccines: we call these the *distrust subset*. The Venn diagram in Fig. 2(right panel) shows how the distrust has now spread across a broad range of topics. This new breadth of topics means that bad-actor-AI content will have a very wide target to aim at in terms of choosing topics to seed widespread unrest. Hence many of these communities and their members could soon get dragged into being part of the bad-actor subsystem.

There is a key feature of the bad-actor dynamics in Fig. 2 that we will use to form our mathematical modeling of bad-actor-AI control strategies in Fig. 4. Specifically, the data show frequent appearances of new links between bad-actor communities, and also infrequent link disappearances—perhaps as a result of action by platform moderators or simply older links dropping off the bottom of the screen feed. This means that clusters of linked nodes (where each node is a bad-actor community) grow over time as links are added, and these clusters occasionally fragment as links are lost. Hence, there is an ongoing dynamical fusion (i.e. coalescence) of nodes into clusters and infrequent fission (i.e. fragmentation) of such clusters (66). Empirical evidence of this was reported recently in Ref. (66). These fusion–fission dynamics mean that any bad-actor community generating content continually with basic AI such as GPT-2 will be able to quickly and widely spread this content while also increasing its connectivity into the vulnerable mainstream.

Figure 3 presents an approach for estimating when bad-actor-AI attacks will likely occur. References (61, 64) showed that (i) a collection of attacking individuals armed with some kind of technology (of which bad actor online communities are a plausible example) typically perform successive advancements (e.g. successful tasks/attacks) with time intervals τ_n such that $\tau_n = \tau_1 n^{-\beta}$, where $n = 1, 2, 3,$ etc., $\beta > 0$, and τ_1 is the initial time interval that forms the intercept on a log–log plot of τ_n vs. n; and that (ii) $\log \tau_1$ and $\log \beta$ show an approximate linear relationship for different real-world realizations of the same system. Figure 3(C and D) shows results for technologically similar automated-algorithm-cyber systems that we will use to form our estimates. These so-called progress curve patterns can be explained by a dynamical version of the Red Queen hypothesis from evolutionary biology in which an agile attacking entity continually adapts to try to maintain a competitive advantage (Fig. 3B (61, 64)). Suppose $x(n)$ is the bad-actor-AI relative advantage following a previous (n'th) successful event, where $x(n)$ follows a general stochastic walk n^{β}, e.g. a partially correlated random walk (61, 64).

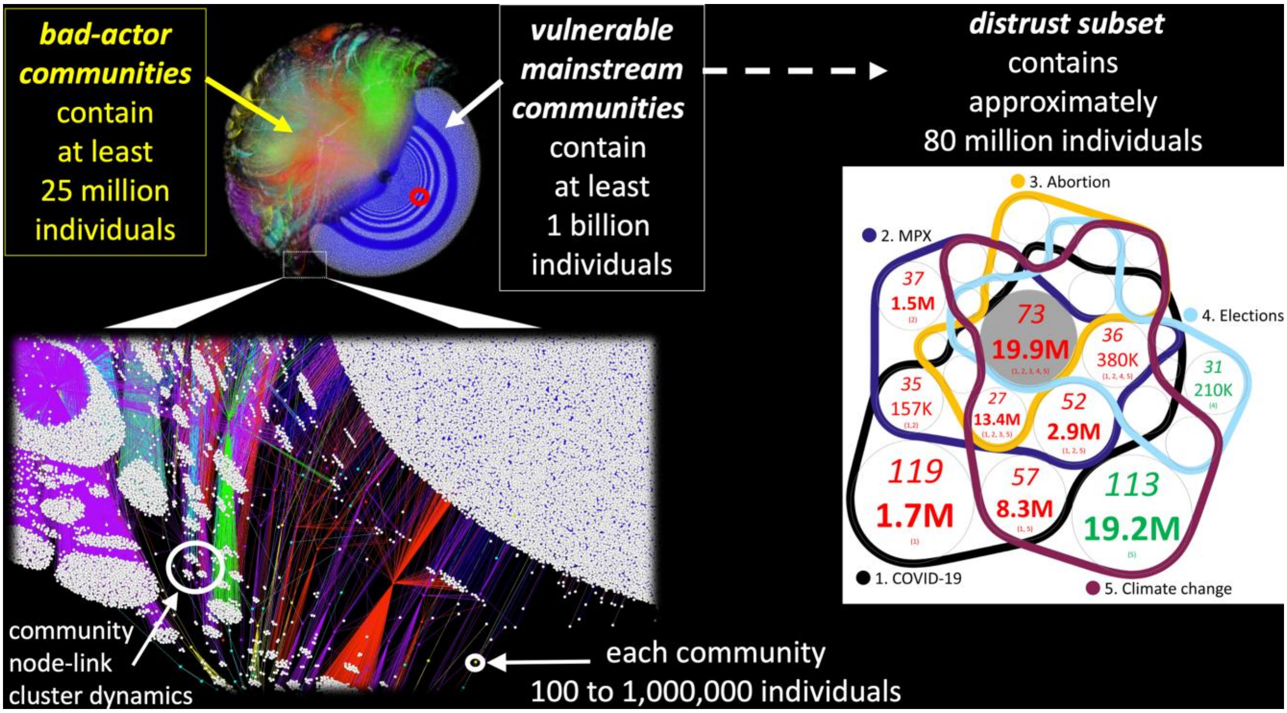

bad-actor communities contain at least 25 million individuals
vulnerable mainstream communities contain 1 billion individuals
distrust subset contains approximately 80 million individuals
3. Abortion
2. MPX
4. Elections
1. COVID-19
5. Climate change
37
1.5M
(2)
73
19.9M
(1, 2, 3, 4, 5)
36
380K
(1, 2, 4, 5)
31
210K
(4)
35
157K
(1, 2)
27
13.4M
(1, 2, 3, 5)
52
2.9M
(1, 2, 5)
119
1.7M
(1)
57
8.3M
(1, 5)
113
19.2M
(5)
community node-link cluster dynamics
each community 100 to 1,000,000 individuals

Fig. 2. This shows where bad-actor-AI activity will likely happen, i.e. across 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 (60) which is spontaneous, hence sets of communities (nodes) appear closer together when they share more links. Different colors correspond to different platforms (see Fig. S1). Small ring shows 2023 Texas shooter's YouTube community as illustration. Ordered circles shows successive sets of white nodes with 1, 2, 3, etc. links from 4Chan hence they experience a net spring force toward the core that is 1, 2, 3, etc. times as strong, so they will be roughly 1, 2, 3, etc. times more likely to receive future bad-actor-AI content and influence. Right panel shows Venn diagram of the topics discussed within the distrust subset (see text and Section S4 for fuller explanation). 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 antivaccination; green if majority is neutral on vaccines. Only regions with >3% of total communities are labeled. Antivaccination 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.

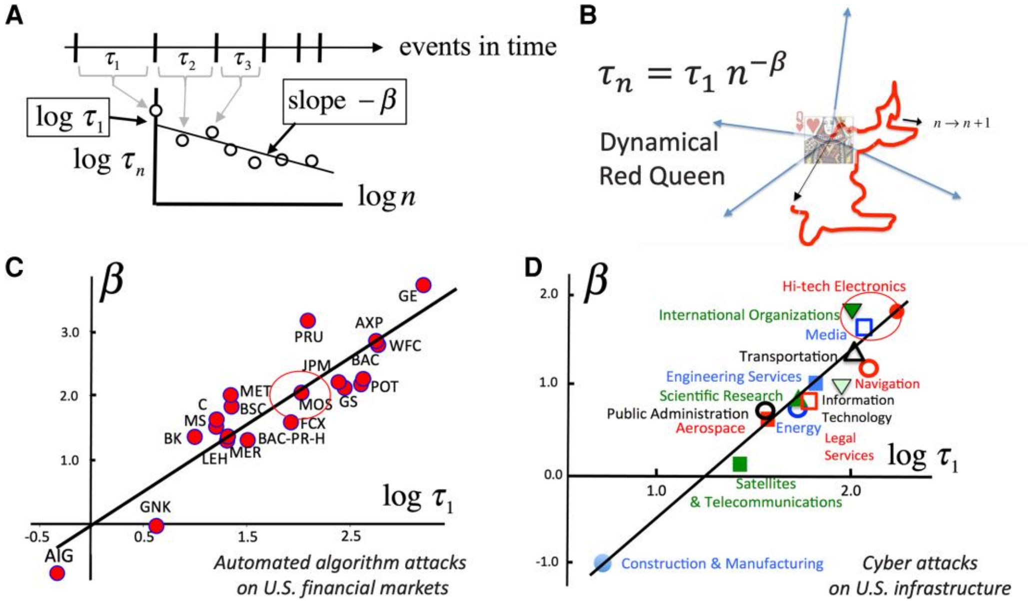
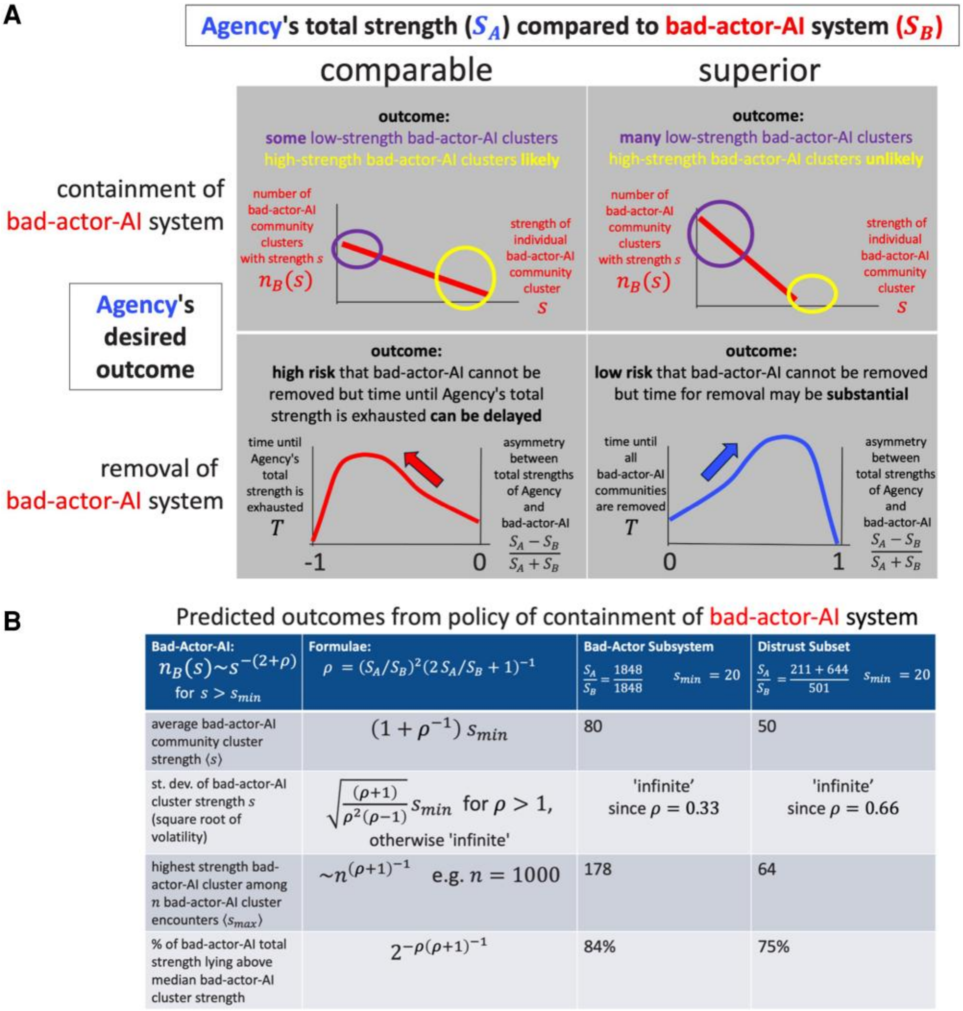


Fig. 3. This offers a prediction of when bad-actor-AI activity will likely happen based on previously observed patterns in similar systems. Specifically, it shows progress curves (61) for the timing of automated algorithm attacks on US financial markets (61, 62) and cyber attacks against US infrastructure (61, 63) adapted by permission from an earlier pilot study by one of us (N.F. Johnson). A) Progress curve predicts successive time intervals between attacks by a general bad actor (61). B) Form of progress curves is explained by bad-actor's advantage following a generalized stochastic walk, e.g. a partially correlated random walk, which is a dynamical generalization of the well-known Red Queen hypothesis from evolutionary biology (see text (61, 64)). Red rings show estimates used for prediction. Overall, this figure provides a quantitative basis for estimating timings of future bad-actor-AI attacks, and hence providing an answer to the question of when bad-actor-AI attacks will likely occur.


A
Agency's total strength (S_A) compared to bad-actor-AI system (S_B)
comparable
superior
containment of bad-actor-AI system
Agency's desired outcome
removal of bad-actor-AI system
outcome:
some low-strength bad-actor-AI clusters
high-strength bad-actor-AI clusters likely
outcome:
many low-strength bad-actor-AI clusters
high-strength bad-actor-AI clusters unlikely
number of bad-actor-AI community clusters with strength s
$n_B(s)$
strength of individual bad-actor-AI community cluster
s
outcome:
high risk that bad-actor-AI cannot be removed but time until Agency's total strength is exhausted can be delayed
outcome:
low risk that bad-actor-AI cannot be removed but time for removal may be substantial
time until Agency's total strength is exhausted
T
time until all bad-actor-AI communities are removed
T
asymmetry between total strengths of Agency and bad-actor-AI
$\frac{S_A - S_B}{S_A + S_B}$
-1
0
0
1
B

Predicted outcomes from policy of containment of bad-actor-AI system

Bad-Actor-AI: $n_B(s) \sim s^{-(2+\rho)}$ for $s > s_{min}$	Formulae: $\rho = (S_A/S_B)^2(2S_A/S_B + 1)^{-1}$	Bad-Actor Subsystem $\frac{S_A}{S_B} = \frac{1848}{1848}$ $s_{min} = 20$	Distrust Subset $\frac{S_A}{S_B} = \frac{211+644}{501}$ $s_{min} = 20$
average bad-actor-AI community cluster strength $\langle s \rangle$	$(1+\rho^{-1})\, s_{min}$	80	50
st. dev. of bad-actor-AI cluster strength s (square root of volatility)	$\sqrt{\frac{(\rho+1)}{\rho^2(\rho-1)}}\, s_{min}$ for $\rho > 1$, otherwise 'infinite'	'infinite' since $\rho = 0.33$	'infinite' since $\rho = 0.66$
highest strength bad-actor-AI cluster among n bad-actor-AI cluster encounters $\langle s_{max} \rangle$	$\sim n^{(\rho+1)^{-1}}$ e.g. $n = 1000$	178	64
% of bad-actor-AI total strength lying above median bad-actor-AI cluster strength	$2^{-\rho(\rho+1)^{-1}}$	84%	75%

Fig. 4. Policy Matrix and Risk Chart for dealing with bad-actor-AI. A) Policy Matrix shows the calculated outcomes for the two policies from Eq. 1 (containment, top row) and Eq. 2 (removal, bottom row). Columns show how outcome changes according to A and B's initial relative strengths (sizes, e.g. number of communities) S_A and S_B. Bottom row: results plotted as a function of asymmetry between initial strengths for A and B for fixed initial total ($S_A + S_B$). B) Risk chart with plug-and-play formulae that predict key outcome risk measures for containment policy (top row in a)). Equations adapted from Ref. (65). Right two columns give two examples using empirical inputs from Fig. 2 and with $s_{min} = 20$ taken from simulations. Because the volatility risk measure is technically infinite for both (second row), A should expect extreme fluctuations in the strength (size) of bad-actor-AI (B) clusters for both systems.

Taking the instantaneous rate of bad-actor-AI successful events as proportional to $x(n)$ and hence n^β, then the time interval $\tau_n = \tau_1 n^{-\beta}$ which yields the progress-curve pattern. Bad-actor-AI activity is too new to have any reliable event data—however, since we are focused on a hi-tech/media/organization setting, we choose estimates $\log \beta \approx 2$ and $\log \tau_1 \approx 2$ from Fig. 3). In addition to coming from a sociotechnical proxy system, these estimates are consistent with the average Fig. 3C values, and also happen to be consistent with the empirical time interval between ChatGPT's initial launch and the arrival of the subsequent wave of variants in 2023 (i.e. crudely $\tau_1 \approx 100$ days hence $\log \tau_1 \approx 2$).

Though obviously crude, we can use these estimates to calculate when the time interval $\tau_n \to 1$ in the progress-curve equation $\tau_n = \tau_1 n^{-\beta}$. The result is that bad-actor-AI attacks are predicted to occur almost daily by mid 2024—in time for the run up to the United States and other global elections. These estimates can be improved using the progress–curve equation $\tau_n = \tau_1 n^{-\beta}$ as actual events occur and hence τ_1, τ_2, etc. become known.

Figure 4 uses a mathematical description of the empirical fusion–fission dynamics (Fig. 2), to examine how the bad-actor-subsystem armed with AI (labeled B) can be controlled by an incumbent agency (labeled A, e.g. pro-X communities). For simplicity, we do not want to assume A has any special powers, hence we take A as undergoing similar dynamics to B, though this can be generalized, and we assume A can only engage clusters of B communities when it finds them. We take B's total strength S_B as the total number of bad-actor-AI communities, and similarly for A, but it could in principle be taken as some more abstract measure. Hence the dynamics involve two populations A (an agency) and B (bad-actor-AI subsystem) with an initial total number of nodes S_A and S_B (number of bad-actor-AI communities) that can aggregate into clusters with their own type; and some form of destructive interaction when A and B clusters meet. We refer to Sections S5 and S6 for full mathematical derivations and demonstration of the good agreement with numerical simulations, as well as Refs. (75, 76).

Consider first the less ambitious policy in which the relevant agency (A) aims to simply contain the bad-actor-AI (B) (Fig. 4A, top row). In this case, we take an interaction between A and a B cluster as simply fragmenting the smaller cluster—which is an easier proposition for A than the alternative of entirely eliminating B

clusters. If $S_A > S_B$, the mathematics in Section S5 show that *A* will be successful in containing *B*—i.e. *A* will be able to control the distribution of *B* clusters' strengths (sizes)—because on average an *A* cluster finding a *B* cluster (i.e. a cluster of *B* communities) will tend to be larger and hence stronger than the *B* cluster if $S_A > S_B$. Hence, it can inactivate the *B* cluster's links by eliminating them from the feed or banning these specific hyperlink connections. This means that the *B* cluster effectively fragments into unlinked *B* communities. In the steady-state (Section S5), the number of *B* clusters with strength *s* will become

$$n_B(s) = Cs^{-\left[2+(S_A/S_B)^2(2S_A/S_B+1)^{-1}\right]} \qquad (1)$$

for $s > s_{\min}$, where *C* is a normalization constant. As S_A/S_B decreases toward unity, the distribution's slope decreases (i.e. the magnitude of the power-law exponent in Eq. 1 decreases) because *A* becomes less able to repartition *B*'s total strength into smaller clusters of communities. This has an important consequence which can be quantified as follows using the volatility risk measure in Fig. 4b. When S_A is less than $(1+\sqrt{2})S_B$ (i.e. $S_B \le S_A \le 2.4S_B$), the standard deviation of the *B* clusters' strength *s* becomes technically infinite since the power-law exponent in Eq. 1 will be <3 (e.g. the power-law exponent is 2.33 if $S_A = S_B$). This means there will be extreme fluctuations in the strength (size) of *B* clusters and hence very strong (i.e. very large) *B* clusters can appear at any time. Even when one very large *B* cluster gets broken up, others will soon build and could become even bigger. But if S_A is greater than $(1+\sqrt{2})S_B$ (i.e. $S_A > 2.4S_B$), the power-law exponent in Eq. 1 becomes >3 and hence the standard deviation in *B*'s cluster strengths becomes finite—hence the chances of very large *B* clusters appearing tends to zero. In Fig. 4b, we also provide other relevant outcomes/risk measures and their estimates for this containment policy calculated from Eq. 1 with empirical sizes S_A and S_B (i.e. number of nodes) estimated from Fig. 2, since these could be used by a relevant agency *A* in the future.

The more ambitious policy of the agency (*A*) aiming to completely remove the bad-actor-AI (*B*) is considered in Fig. 4a bottom row. Now, any interaction between *A* and a *B* cluster leads to removal of the smaller cluster. This means that when $S_A > S_B$, the on-average stronger agency (*A*) cluster finding a *B* cluster will remove it, e.g. it bans all the *B* cluster's communities. This is more challenging for *A* and may be more widely criticized as censorship. The time for *A* to completely remove *B* given initial strengths S_A, S_B, becomes:

$$T = 2S_B + \frac{1}{2}(S_A - S_B)\ln\left(S_B(S_A - S_B)/S_A\right). \qquad (2)$$

This reveals a further downside to the goal of complete *B* removal: as A's strength increases, the *B* clusters become less strong (i.e. smaller) on average and hence less noticeable to *A*. This creates a rise and peak in the time needed *T* since it takes increasingly long for *A* to find *B* clusters. If on the other hand *B* is stronger than *A* ($S_B > S_A$), then *B* cannot be removed—but the slight silver lining for *A* is that this large *T* means an extensive time until *A*'s strength is exhausted. Overall, containment may seem the better choice—but Fig. 4 provides a quantitative way for different agencies to come their own conclusions.

Since nobody can predict exactly what will happen with future bad-actor-AI given the rapid pace of technology and changing online landscape, the predictions in this article are strictly speaking speculative. But they are each quantitative and testable—and also generalizable—and hence provide a concrete starting point for strengthening bad-actor-AI policy discussions. We realize that many features of our analysis could be extended and improved: for example, what happens if future AI can predict the cluster dynamics (ChatGPT currently cannot) and hence bad-actor-AI community clusters outwit the containment mechanism? They could also use new decentralized or block-chain platforms as perpetual GPT reactor cores that generate unstoppable streams of bad-actor content. Our label "bad actor" should be subclassified (e.g. anti-Semitic vs. antiwomen) as should "vulnerable mainstream community." We should also account for links from the vulnerable mainstream communities back into bad-actor communities. Going forward, our predictions can be adjusted as bad-actor-AI capabilities evolve.

Conclusion

This study presented a fresh approach to analyzing, modeling, and hence formatting policies for the proliferation of bad-actor-AI activities in online spaces with a particular focus on social media. Our integration of empirical data with dynamical systems modeling sheds light on current and potential trajectories of bad-actor-AI activity. The findings clarify some of the significant challenges to platform moderators and legislation, posed by bad actors with just basic AI versions. Our study also highlights the frequently overlooked role of smaller platforms in providing links that help bind together the larger ecosystem; and it suggests the increasing risk of AI misuse in sync with future global events like elections. The proposed Policy Matrix and mathematical tools for evaluating containment vs. removal policies, offer potentially valuable insights into managing the threats. Although this work establishes a foundational framework for addressing bad-actor-AI risks, it also signals the necessity for continuous research in this field, especially considering the rapid advancement of AI technologies and the ever-changing landscape of the online community ecosystem at scale.

Acknowledgments

The authors are very grateful to Minzhang Zheng and Alex Dixon for help with figures and prior analyses and discussions.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

N.F.J. is supported by US Air Force Office of Scientific Research awards FA9550-20-1-0382 and FA9550-20-1-0383, and by The John Templeton Foundation.

Author Contributions

N.F.J. supervised the project, performed some of the analysis, and wrote the article drafts. R.S. collected the data, managed databases, and developed software. L.I. analyzed the data and produced images.

Preprints

A preprint of an earlier working version of this article is available at arXiv.2308.00879.

Data Availability

Data that reproduce the figures are available online at the GW Donlab website https://github.com/gwdonlab/data-access. This provides readers with access to the minimum dataset that is necessary to interpret, verify, and extend the research in this article. The code used to prepare the data was standard Python libraries for web crawling (e.g. BeautifulSoup, lxml), quantitative analysis (e.g. pandas, numpy), and data visualization (Plotly), all of which are open and free. Gephi (also free and open-source) was used to produce network visualizations. Mathematica was also used for quantitative analysis and Adobe Illustrator was used to produce the final figures. These are well-known commercial products available through site licenses in many universities. Figure 1 uses code from Wolfram Mathematica and it can be downloaded at writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/. This software allows readers to reproduce and explore further the distribution in Fig. 1b, and to produce text like that in Fig. 1a.

References

1 AI experts predict by 2026, 90% of online content will be generated by artificial intelligence. 2022. International Data Center Authority Press [accessed 2022 Sep 27]. https://idc-a.org/news/industry/AI-Experts-Predict-By-2026-90-Of-Online-Content/127ab0c0-34ba-4c03-8bad-1e4f21923f31.

2 Hoffmann M, Frase H. 2023. Adding structure to AI harm. Center for Security and Emerging Technology Publications. https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/.

3 Milmo D, Hern A. 2023. Elections in UK and US at risk from AI-driven disinformation, say experts. The Guardian [accessed 2023 May 20]. https://www.theguardian.com/technology/2023/may/20/elections-in-uk-and-us-at-risk-from-ai-driven-disinformation-say-experts.

4 Hsu T, Myers SL. 2023. A.I.'s use in elections sets off a scramble for guardrails. The New York Times [accessed 2023 Jun 25]. https://www.nytimes.com/2023/06/25/technology/ai-elections-disinformation-guardrails.html.

5 Euchner T. T. 2023. Coping with the fear of mass shootings. *Siouxland Proud*. https://www.siouxlandproud.com/news/local-news/coping-with-the-fear-of-mass-shootings/.

6 One-Third of U.S. adults say fear of mass shootings prevents them from going to certain places or events. Social Work Today [accessed 2024 Jan 10]. https://www.socialworktoday.com/news/dn_081519.shtml.

7 Strengthened code of practice on disinformation: signatories to identify ways to step up work one year after launch. 2023. European Commission Press Release. https://digital-strategy.ec.europa.eu/en/news/strengthened-code-practice-disinformation-signatories-identify-ways-step-work-one-year-after-launch.

8 The Artificial Intelligence Act. 2024. European Union website [accessed 2024 Jan 10]. https://web.archive.org/web/20230811085634/https://artificialintelligenceact.eu/.

9 Digital Services Act: Commission designates first set of very large online platforms and search engines. 2023. European Commission Press Release. https://digital-strategy.ec.europa.eu/en/news/digital-services-act-commission-designates-first-set-very-large-online-platforms-and-search-engines.

10 Benninger M. 'Fringe' websites radicalized Buffalo shooter, report concludes. 2022. WBNG News website [accessed 2022 Oct 18]. https://www.wbng.com/2022/10/18/fringe-websites-radicalized-buffalo-shooter-report-concludes/.

11 Fringe social media: are you digging deep enough? 2019. SMI Aware website [accessed 2019 Oct 8]. https://web.archive.org/web/20201001222412/https://smiaware.com/blog/fringe-social-media-are-you-digging-deep-enough/.

12 Rodrigo CM, Klar R. 2021. Fringe social networks boosted after mob attack. The Hill [accessed 2021 Jan 12]. https://thehill.com/policy/technology/533919-fringe-social-networks-boosted-after-mob-attack/.

13 Hsu T. 2022. News on fringe social sites draws limited but loyal fans, report finds. The New York Times [accessed 2022 Oct 6]. https://www.nytimes.com/2022/10/06/technology/parler-truth-social-telegram-pew.html.

14 Dewey C. 2022. On fringe social media sites, Buffalo mass shooting becomes rallying call for white supremacists. Buffalo News [accessed 2022 Aug 7]. https://buffalonews.com/news/local/on-fringe-social-media-sites-buffalo-mass-shooting-becomes-rallying-call-for-white-supremacists/article_74a55388-f61b-11ec-812a-97d8f2646d45.html.

15 Scott M. 2022. Fringe social media networks sidestep online content rules. Politico [accessed 2022 Jan 25]. https://www.politico.eu/article/fringe-social-media-telegram-extremism-far-right/.

16 Kupferschmidt K. 2023. Does social media polarize voters? Unprecedented experiments on Facebook users reveal surprises. Science [accessed 2023 Jul 27]. https://www.science.org/content/article/does-social-media-polarize-voters-unprecedented-experiments-facebook-users-reveal.

17 Social Media and Elections. 2023 Special Issue: Science Vol. 381 Issue 6656 [accessed 2023 Jul 28]. https://www.science.org/toc/science/381/6656.

18 Uzogara EE. 2023. Democracy intercepted. *Science*. 381(6656):386–387.

19 González-Bailón S, *et al.* 2023. Asymmetric ideological segregation in exposure to political news on facebook. *Science*. 381(6656):392–398.

20 Guess AM, *et al.* 2023. How do social media feed algorithms affect attitudes and behavior in an election campaign?. *Science*. 381(6656):398–404.

21 Guess AM, *et al.* 2023. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*. 381(6656):404–408.

22 Nyhan B. 2023. Like-minded sources on Facebook are prevalent but not polarizing. *Nature*. 620(7972):137–144.

23 Aut N, Ranaware S, Ghadge S, Jadhav R, Jagtap P. 2023. Social media based hate speech detection using machine learning. *Int J Res Appl Sci Eng Technol*. 11(V):2729–2736. https://www.ijraset.com/best-journal/social-media-based-hate-speech-detection-using-machine-learning.

24 Ollagnier A, Cabrio E, Villata S. 2023. Harnessing bullying traces to enhance bullying participant role identification in multi-party chats. In: The International FLAIRS Conference Proceedings. https://journals.flvc.org/FLAIRS/article/view/133191.

25 Aldreabi E, Lee JM, Blackburn J. 2023. Using deep learning to detect islamophobia on Reddit. In: The International FLAIRS Conference Proceedings. https://journals.flvc.org/FLAIRS/article/view/133324.

26 Morgan M, Kulkarni A. 2023. Platform-agnostic model to detect Sinophobia on social media. In: ACMSE 2023: Proceedings of the 2023 ACM Southeast Conference. New York (NY): Association for Computing Machinery. p. 149–153. doi: 10.1145/3564746.3587024.

27 Beacken G, Trauthig I, Woolley S. 2022. Platforms' efforts to block antisemitic content are falling short. https://www.cigionline.org/articles/platforms-efforts-to-block-anti-semitic-content-are-falling-short/.

28 Cinelli M, *et al.* 2021. Dynamics of online hate and misinformation. *Sci Rep*. 11(1):22083.

29 Chen E, Lerman K, Ferrara E. 2020. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. *JMIR Public Health Surveill*. 6(2): e19273.

30 Gelfand MJ, Harrington JR, Jackson JC. 2017. The strength of social norms across human groups. *Perspect Psychol Sci*. 12(5): 800–809.

31 van der Linden S, Leiserowitz A, Rosenthal S, Maibach E. 2017. Inoculating the public against misinformation about climate change. *Global Chall*. 1(2):1600008. https://onlinelibrary.wiley.com/doi/pdf/10.1002/gch2.201600008.

32 Lewandowsky S, *et al.* 2020. Debunking handbook 2020. Technical report, George Mason University.

33 Lazer DMJ, *et al.* 2018. The science of fake news. *Science*. 359(6380): 1094–1096.

34 Smith R, Cubbon S, Wardle C. 2020. Under the surface: Covid-19 vaccine narratives, misinformation and data deficits on social media. Technical report, First Draft.

35 Semenov A, *et al.* 2019. Exploring social media network landscape of post-Soviet space. *IEEE Access*. 7:411–426.

36 Rao A, Morstatter F, Lerman K. 2022. Partisan asymmetries in exposure to misinformation. *Sci Rep*. 12(1):15671.

37 Wu X-Z, Fennell PG, Percus AG, Lerman K. 2018. Degree correlations amplify the growth of cascades in networks. *Phys Rev E*. 98(2):022321.

38 Roozenbeek J, van der Linden S, Goldberg B, Rathje S, Lewandowsky S. 2022. Psychological inoculation improves resilience against misinformation on social media. *Sci Adv*. 8(34): eabo6254.

39 Biever C. 2023. ChatGPT broke the Turing test—the race is on for new ways to assess AI. *Nature*. 619(7971):686–689.

40 Miller-Idriss C. 2020. Hate in the homeland: the new global far right. Princeton (NJ): Princeton University Press.

41 The haters and conspiracy theorists back on Twitter. BBC News. 2023. https://www.bbc.com/news/technology-64554381.

42 DiResta R. 2018. The digital maginot line. https://www.ribbonfarm.com/2018/11/28/the-digital-maginot-line/.

43 Vesna C-G, Maslo-Čerkić Š. 2023. Hate speech online and the approach of the Council of Europe and the European Union. Technical report, University of Rijeka, Tallinn. https://urn.nsk.hr/urn:nbn:hr:118:377009.

44 House of Commons Home Affairs Committee. 14th Report - Hate crime: abuse, hate and extremism online. Government Report HC 609, House of Commons, London. 2017.

45 Hart R. 2023. White supremacist propaganda hit record levels in 2022, ADL Says. Forbes [accessed 2023 Mar 9]. https://www.forbes.com/sites/roberthart/2023/03/09/white-supremacist-propaganda-hit-record-levels-in-2022-adl-says/.

46 Online hate and harassment: the American Experience. Technical report, ADL Center for Technology & Society. https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2023.

47 Eisenstat Y. 2023. Hate is surging online—and social media companies are in denial. Congress can help protect users. https://thehill.com/opinion/congress-blog/4085909-hate-is-surging-online-and-social-media-companies-are-in-denial-congress-can-help-protect-users/.

48 Nelson DJ. 2023. UN warns of AI-generated deepfakes fueling hate and misinformation online. https://decrypt.co/144281/un-united-nations-ai-deepfakes-hate-misinformation.

49 United Nations. Common agenda policy brief: information integrity on digital platforms. Technical report. 2023. https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf.

50 Brown R, Livingston L. 2018. A new approach to assessing the role of technology in spurring and mitigating conflict: evidence from research and practice. https://jia.sipa.columbia.edu/news/new-approach-assessing-role-technology-spurring-and-mitigating-conflict-evidence-research-and.

51 Starbird K. 2019. Disinformation's spread: bots, trolls and all of us. *Nature*. 571(7766):449–449.

52 Lamensch M. To eliminate violence against women, we must take the fight to online spaces. https://www.cigionline.org/articles/to-eliminate-violence-against-women-we-must-take-the-fight-to-online-spaces/.

53 Crawford A, Smith T. 2023. Illegal trade in AI child sex abuse images exposed. BBC News. https://www.bbc.com/news/uk-65932372.

54 Cosoleto T. 2023. Surge in young children being targeted by cyber bullies. https://thewest.com.au/news/social/surge-in-young-children-being-targeted-by-cyber-bullies-c-11223220.

55 Gill P, Corner E, Jarvis L, Macdonald S, Chen T. 2015. Lone actor terrorist use of the Internet and behavioural correlates. In: Jarvis L, Macdonald S, Chen T, editors. Terrorism online: politics, law, technology and unconventional violence. Chapter 2. Oxford (UK): Routledge.

56 Douek E. 2022. Content moderation as systems thinking. *Harv Law Rev*. 136(2):526–607. https://harvardlawreview.org/print/vol-136/content-moderation-as-systems-thinking/.

57 Dynamic online networks laboratory. Literature review. Technical report. bpb-us-e1.wpmucdn.com/blogs.gwu.edu/dist/5/3446/files/2022/10/lit_review.pdf.

58 Jusup M, *et al.* 2022. Social physics. *Phys Rep*. 948:1–148.

59 DisinfoDocket. https://www.disinfodocket.com/.

60 Jacomy M, Venturini T, Heymann S, Bastian M. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*. 9(6):e98679.

61 Johnson NF, *et al.* 2013. Simple mathematical law benchmarks human confrontations. *Sci Rep*. 3(1):3463.

62 Data from NANEX. https://www.nxcoredata.com/.

63 Data from MANDIANT. https://www.mandiant.com/.

64 Johnson N, *et al.* 2011. Pattern in escalations in insurgent and terrorist activity. *Science*. 333(6038):81–84.

65 Newman M. 2005. Power laws, Pareto distributions and Zipf's law. *Contemp Phys*. 46(5):323–351.

66 Manrique PD, *et al.* 2023. Shockwavelike behavior across social media. *Phys Rev Lett*. 130(23):237401.

67 Lupu Y, *et al.* 2023. Offline events and online hate. *PLoS One*. 18(1): e0278511.

68 Velásquez N, *et al.* 2021. Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms. *Sci Rep*. 11(1):11549.

69 Ammari T, Schoenebeck S. 2016. 'Thanks for your interest in our Facebook group, but it's only for dads': social roles of stay-at-home dads. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16. New York (NY): Association for Computing Machinery. p. 1363–1375.

70 Moon RY, Mathews A, Oden R, Carlin R. 2019. Mothers' perceptions of the internet and social media as sources of parenting and health information: qualitative study. *J Med Internet Res*. 21(7):e14289.

71 Laws R, *et al.* 2019. Differences between mothers and fathers of young children in their use of the internet to support healthy family lifestyle behaviors: cross-sectional study. *J Med Internet Res*. 21(1):e11454.

72 Madhusoodanan J. 2022. Safe space: online groups lift up women in tech. *Nature*. 611(7937):839–841.

73 Kilcher Y. 2023. GPT-4chan model card. https://www.ykilcher.com/gpt-4chan-model-card.

74 Leahy R, Restrepo NJ, Sear R, Johnson NF. 2022. Connectivity between Russian information sources and extremist communities across social media platforms. *Front Polit Sci*. 4:885362. https://doi.org/10.3389/fpos.2022.885362.

75 Dixon A, Zhao Z, Bohorquez JC, Denney R, Johnson N. 2010. Statistical physics and modern human warfare. In: Naldi G, Pareschi L, Toscani G, editors. *Mathematical modeling of collective behavior in socio-economic and life sciences, modeling and simulation in science, engineering and technology*. Boston: Birkhäuser. p. 365–396.

76 Zhao Z, Bohorquez JC, Dixon A, Johnson NF. 2009. Anomalously slow attrition times for asymmetric populations with internal group dynamics. *Phys Rev Lett*. 103(14):148701.