Journal of Quantitative Description: Digital Media 4(2024), 1–36 10.51685/jqd.2024.icwsm.7

Anatomy of an AI-powered malicious social botnet

KAI-CHENG YANG

Network Science Institute, Northeastern University, USA

Observatory on Social Media, Indiana University, USA

FILIPPO MENCZER

Observatory on Social Media, Indiana University, USA

Large language models (LLMs) exhibit impressive capabilities in generating realistic text across diverse subjects. Concerns have been raised that they could be utilized to produce fake content with a deceptive intention, although evidence thus far remains anecdotal. This paper presents a case study about a Twitter botnet that appears to employ ChatGPT to generate human-like content. Through heuristics, we identify 1,140 accounts and validate them via manual annotation. These accounts form a dense cluster of fake personas that exhibit similar behaviors, including posting machine-generated content and stolen images, and engage with each other through replies and retweets. ChatGPT-generated content promotes suspicious websites and spreads harmful comments. While the accounts in the AI botnet can be detected through their coordination patterns, current state-of-the-art LLM content classifiers fail to discriminate between them and human accounts in the wild. These findings highlight the threats posed by AI-enabled social bots.

Keywords: social bot, AI, large language model, coordinated inauthentic operation

Kai-Cheng Yang: yang3kc@gmail.com
Date submitted: 2023-10-06

Large language models (LLMs) can generate human-like text (Jakesch et al., 2023) and excel in various natural language processing tasks, including sentiment analysis and text summarization (Ye et al., 2023; Qin et al., 2023). Such capabilities have opened up numerous potential applications (Bahrini et al., 2023), such as enhancing education (Kasneci et al., 2023) and providing responses to medical questions (Ayers et al., 2023). Consequently, these tools have gained significant traction in the market. For instance, ChatGPT, a prominent LLM, amassed over 100 million users in only two months.[1] Industry giants like Microsoft have integrated ChatGPT into their products,[2] and Google swiftly followed suit by offering similar solutions.[3]

At the same time, researchers have warned about the potential misuse of such technologies to generate mis/disinformation and harmful content (Bommasani et al., 2021; Yamin et al., 2021; Guembe et al., 2022; Goldstein et al., 2023). Recent studies demonstrate that models like the GPT series are capable of producing news articles indistinguishable from human-generated ones (Kreps et al., 2022; Jakesch et al., 2023) and vast amounts of compelling mis/disinformation with little human involvement (Buchanan et al., 2021; Spitale et al., 2023). Detecting such content poses challenges for existing automated detection models (Zhou et al., 2023). LLMs can also be leveraged to scale up personalized attacks, such as spear phishing content (Hazell, 2023). The widespread availability of powerful language models will also significantly reduce the costs associated with content generation and lower the technical proficiency required to conduct such operations (Brundage et al., 2018; Weidinger et al., 2022). However, existing studies mainly consist of predictions and lab experiments. To date, evidence of LLMs being deployed in the field for malicious purposes remains largely anecdotal.

In this paper, we present a case study about a Twitter botnet that appears

[1] reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01 (Accessed May 2023)
[2] blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web (Accessed May 2023)
[3] blog.google/products/search/generative-ai-search (Accessed May 2023)

to use ChatGPT to generate harmful content. Social bots are social media accounts controlled in part by software and have been around for many years (Ferrara et al., 2016). They were found to distort online conversations and spread misinformation in various contexts, from elections to public health crises (Shao et al., 2018; Ferrara et al., 2020; Jamison et al., 2019; Marlow et al., 2020). Traditional social bots often follow pre-defined instructions to perform simplistic tasks, such as spamming (Yang et al., 2019), following others, and amplifying certain narratives (Keller et al., 2020). They typically lack the intelligence to create realistic personas, post convincing content, or carry out natural conversations with other accounts automatically (Assenmacher et al., 2020). However, the recent advancements in and wide adoption of LLMs have completely transformed this landscape. Adversarial actors can now easily leverage language models to significantly enhance the capabilities of bots across all dimensions.

The bot accounts in our analytical sample were identified by self-revealing tweets they posted by accident. A combination of heuristics and manual annotation yields 1,140 accounts in what we call the "fox8" botnet. An in-depth analysis of behaviors displayed by the accounts in this botnet shows that they form a dense social network by following each other. They post machine-generated content and steal selfies to create fake personas. They also frequently interact with each other through retweets and replies. A closer look at the self-revealing tweets suggests that the ChatGPT-generated content aims to promote suspicious websites and spread harmful comments. We apply state-of-the-art LLM-content detectors and find they cannot effectively distinguish between human and LLM-powered bots in the wild.

Our work unveils the emergence of LLM-powered social bots and highlights the threats they pose. By focusing on a real-world botnet, we provide valuable insights into how LLMs are leveraged by adversarial actors in the field. Given the rapid advancements in AI technologies, we anticipate the proliferation of more advanced bot accounts across social media, serving diverse purposes. We hope to raise public awareness about this issue and share the botnet data to allow the research community to investigate further.

Related work

LLM-powered cyber threats

Machine-generated content has long been implicated in cyber-social threats, such as phishing, mis/disinformation, and harmful content (Crothers et al., 2023). These threats have been further exacerbated by LLMs for two main reasons. First, LLMs beat traditional text generation methods in producing human-like text (Jakesch et al., 2023; Guo et al., 2023; Clark et al., 2021). This enhancement enables them to craft compelling and personalized content for previously unseen attacks (Buchanan et al., 2021; Spitale et al., 2023; Hazell, 2023).

Second, powerful LLMs have become readily accessible, affordable, and user-friendly. For instance, OpenAI provides API access to their models, enabling users to generate large volumes of content at a nominal expense.[4] Interaction with most LLMs happens via human language prompts (Liu et al., 2023), enabling even those without technical skills to harness the models' capabilities. Users can also acquire knowledge on API queries and LLM prompting directly from the models themselves (Hazell, 2023). The availability of open-source LLMs offers technical users greater flexibility to train, customize, and implement models according to their needs (Yang et al., 2023).

Therefore, LLMs have the potential to reshape the landscape of cyber-social security dramatically, a concern shared by many researchers (Bommasani et al., 2021; Yamin et al., 2021; Guembe et al., 2022; Goldstein et al., 2023; Kucharavy et al., 2023). However, empirical evidence of such abuse in the field is rare. One example comes from Hanley et al., who analyze machine-generated text in news media outlets and find a significant surge in such content after the release of ChatGPT, particularly on low-credibility websites (Hanley and Durumeric, 2023). Our research contributes to this growing body of work, focusing on a botnet that exploits ChatGPT for harmful

[4]openai.com/blog/introducing-chatgpt-and-whisper-apis (Accessed May 2023)

activities.

LLM-generated content detection

The potential misuse of LLMs necessitates the development of reliable methods to detect LLM-generated content (Crothers et al., 2023). Existing strategies can be broadly classified into black-box and white-box approaches (Tang et al., 2024). Black-box detection methods are often framed as binary classification problems where classifiers are trained on texts generated by humans and machines. The goal is to identify the characteristics of machine-generated content, such as statistical anomalies (Gehrmann et al., 2019; Mitchell et al., 2023) and linguistic patterns (Guo et al., 2023; Fröhling and Zubiaga, 2021). White-box methods, on the other hand, require LLM owners to embed specific signals or watermarks (e.g., altered word frequencies) into generated content for subsequent identification (Kirchenbauer et al., 2023; Zhao et al., 2023).

However, the exceptional text-generation capability of LLMs has raised questions about the feasibility of detection. For example, Sadasivan et al. argue that black-box detection might be unachievable when LLMs can produce text completely indistinguishable from human-generated content (Sadasivan et al., 2023). Chakraborty et al. show that detection is theoretically possible if machine- and human-generated content distributions differ, but as LLMs advance, the sample size required for detection increases (Chakraborty et al., 2023). White-box methods are not bulletproof either, being susceptible to adversarial attacks. Recent studies suggest that text paraphrasing significantly reduces detection accuracy (Sadasivan et al., 2023; Krishna et al., 2024). Furthermore, this approach might not apply to open-source LLMs, as malicious actors could intentionally remove the embedded watermarks (Tang et al., 2024). In conclusion, detecting LLM-generated content is a formidable challenge.

Social bot detection

Different from chatbots such as ChatGPT, which interact with users solely through text, social media bots display profiles and engage with others through various means, including following, liking, and retweeting. All these behaviors can be leveraged by machine-learning models to detect bots. Researchers commonly adopt the supervised approach where examples of bot and human accounts are collected for training classifiers (Yang et al., 2019; Sayyadiharikandeh et al., 2020). Various characteristics, ranging from account metadata to social network structure, are then considered during the process (Ferrara et al., 2016; Varol et al., 2017). Content posted by bots also provides essential clues (Kudugunta and Ferrara, 2018; Heidari and Jones, 2020).

As some bots evolve to mimic human profiles and behaviors better, their activities can only be detected during orchestrated operations (Cresci, 2020). This unsupervised approach typically requires calculating the similarity of different accounts and subsequently clustering them into groups (Pacheco et al., 2021). Key signals include temporal activities (Chavoshi et al., 2016; Keller et al., 2017), common retweets (Nizzoli et al., 2021), and URLs shared in tweets (Pacheco et al., 2021; Giglietto et al., 2020). Definitions of the similarity measure vary across studies. Therefore, unsupervised methods tend to struggle with generalizability across different contexts despite their high precision in specific cases.

Bots supercharged by LLMs

LLMs also have the potential to enhance the capabilities of social bots, akin to previously mentioned cyber-social threats (Grimme et al., 2022; Ferrara, 2023). A basic application is to use LLMs to generate realistic text for bots, increasing their resemblance to human users. To develop effective detection methods for this kind of bot, Kumarage et al. construct an in-house dataset employing GPT-2 to generate text and compare it with human-generated content (Kumarage et al., 2023). However, empirical investigations into social bots leveraging machine-generated text

are limited. A noteworthy exception is the work by Fagni et al., who identify 23 self-disclosed bot accounts on Twitter and share the dataset (Fagni et al., 2021). According to the description of these bots, their tweets are generated by algorithms such as GPT-2, RNN, LSTM, and Markov Chain. Subsequent studies have built upon this dataset to explore various strategies for bot detection based on content (Saravani et al., 2021; Gambini et al., 2022; Tourille et al., 2022). Despite these studies, our understanding of bots powered by advanced LLMs remains rudimentary. The present study contributes to this line of research by analyzing a more up-to-date botnet that utilizes state-of-the-art AI models.

Identification of the fox8 botnet

As mentioned above, the fox8 botnet was identified through self-revealing tweets posted by these accounts accidentally. To prevent the generation of undesirable content, proprietary LLMs often have safeguards implanted through a technique called reinforcement learning from human feedback (Ouyang et al., 2022). ChatGPT models, for example, are instructed to refuse to respond to any questions[5] that go against OpenAI's usage policies.[6] Violations include harmful content, disinformation, and tailored financial advice. Upon a violation, the models respond with a standardized message asserting their identity as AI language models and their inability to comply (see Table 1 for examples). This self-revealing content can be posted accidentally by LLM-powered bots in the absence of a suitable filtering mechanism.

Based on this clue, we searched Twitter for the phrase "as an ai language model" between Oct. 1, 2022, and Apr. 23, 2023, using the historical search endpoint of Twitter's V2 API. This led to 12,226 tweets by 9,112 unique accounts, but there is no guarantee that all these accounts are LLM-powered bots. Therefore, we selected a sample of 100 accounts for manual verification, discovering that 76% are likely humans posting or retweeting ChatGPT outputs, while the remaining accounts are

[5]openai.com/blog/how-should-ai-systems-behave (Accessed May 2023)
[6]openai.com/policies/usage-policies (Accessed May 2023)

likely bots using LLMs for content generation. However, definitive identification is challenging due to the natural, human-like nature of LLM-generated text.

During the annotation process, we noticed recurring patterns among some bot-like accounts. Specifically, they consistently link to three suspicious websites: fox8.news[7] (distinct from the legitimate news outlet, fox8.com), cryptnomics.org,[8] and globaleconomics.news.[9] Consequently, we extracted all 1,140 accounts linking to any of these websites for further investigation and found strong indications of their origin from the same botnet, likely employing ChatGPT for content creation (see evidence in the following sections). Therefore, we dub it the "fox8" botnet and focus on it in this study.

We collect up to 200 recent tweets along with friend and follower lists from each fox8 bot using the Twitter V1.1 API for further investigation. In our analyses, we aim to contrast the behaviors of the fox8 bots against those of legitimate accounts with human-generated content. To do so, we turn to pre-existing datasets employed for training social bot detectors (Yang et al., 2022). Specifically, we utilize four datasets: `botometer-feedback` (Yang et al., 2019), `gilani-17` (Gilani et al., 2017), `midterm-2018` (Yang et al., 2020), and `varol-icwsm` (Varol et al., 2017), randomly selecting 285 human accounts from each. The accounts in these datasets were annotated by humans. Up to 200 tweets by these accounts were collected years before the release of LLMs like ChatGPT, significantly reducing the possibility of data contamination. Combining the 1,140 bot accounts with the 1,140 human ones results in our benchmark dataset: the `fox8-23` dataset, which is publicly available at github.com/osome-iu/AIBot_fox8.

Characterizations

In this section, we characterize the fox8 bots to reveal their behavioral patterns.

[7] web.archive.org/web/20230401111956/https://fox8.news
[8] web.archive.org/web/20230401190830/https://cryptnomics.org
[9] web.archive.org/web/20230401111503/https://globaleconomics.news

Profiles

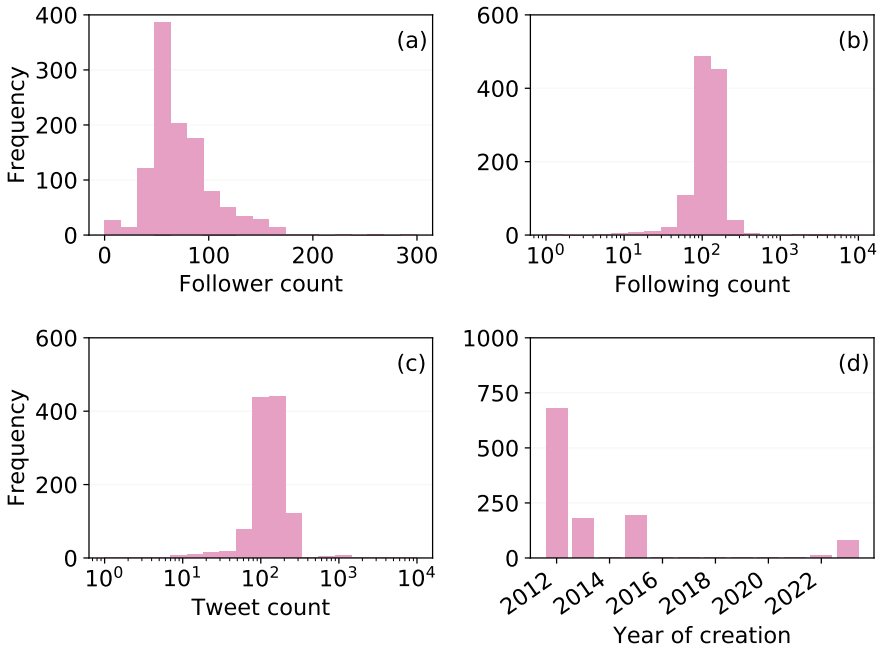

Figure 1. Profile characteristics of the fox8 bots (N=1,140).
Note. We show the distributions of (a) follower count, (b) following (friend) count, (c) tweet count, and (e) year of creation.

Let us start with fox8 profiles and show the distributions of their follower/following count, tweet count, and creation year in Figure 1. These bots have 74.0 (SD=36.7) followers, 140.4 (SD=236.6) friends, and 149.6 (SD 178.8) tweets on average. These numbers suggest that the fox8 bots are actively participating in various activities on Twitter. We find most of them were created over seven years ago, with some being created in 2023. Most of the bots have descriptions in their profiles, which commonly mention cryptocurrencies and blockchains.

Social networks

Let us analyze the social networks of the fox8 bots. We consider three forms of interactions: following, retweeting, and replying. Quotes are ignored due to their

rareness. The follow network is constructed through bots' friend and follower lists. The retweet and reply networks are inferred based on their recent tweets. Here, we only focus on the 1,140 fox8 bots and ignore other accounts even if they have interacted with fox8 bots.

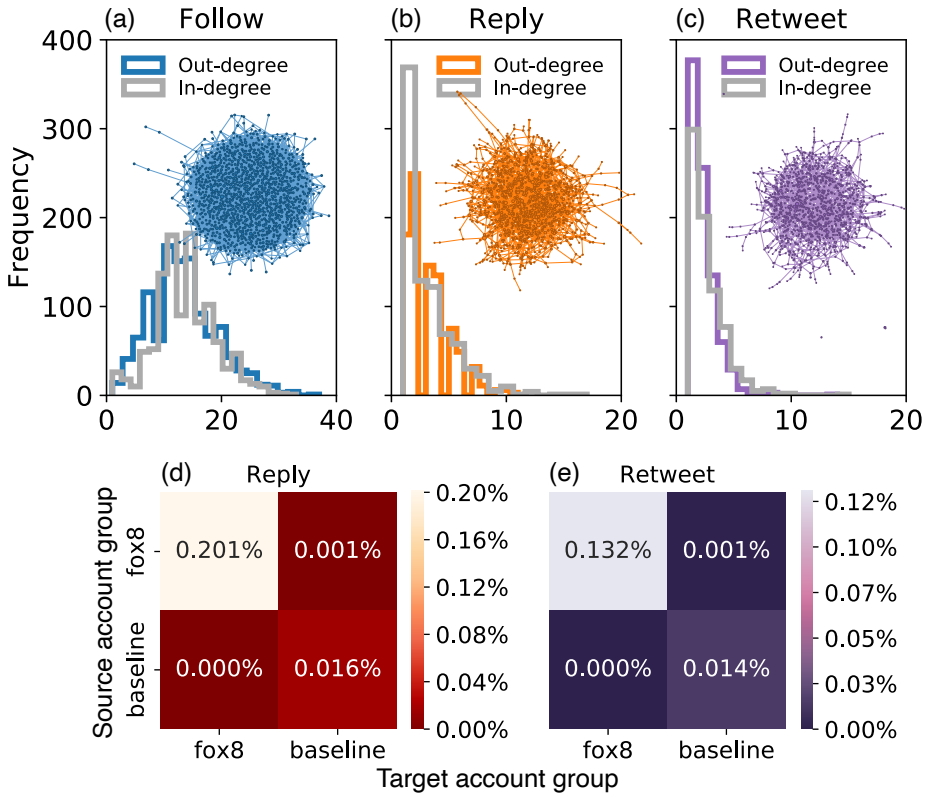

Figure 2. Social networks of the fox8 bots.

Note. (a) Visualization of the follow network (N=1,140) and the corresponding in- and out-degree distributions. (b) Same as (a) but for the reply network (N=1,036). (c) Same as (a) but for the retweet network (N=1,058). (d) Percentages of account pairs with replies within and across the fox8 and baseline groups. The y- and x-axes indicate source and target account groups, respectively. (e) Same as (d) but for retweets.

We visualize the follow network in Figure 2(a), which turns out to be very dense: it has an in-degree of 13.7 (SD=5.2) and an out-degree of 13.4 (SD=5.8) on average. The near-identical distributions concentrated around mean values suggest that the following behaviors of the fox8 bots are engineered rather than organic —

empirical distributions of follower counts tend to be significantly broader and more skewed toward lower values (Conover et al., 2012). Similarly, we show the reply and retweet networks in Figure 2(b,c). The reply network is much sparser, with an average in-degree of 3.4 (SD=2.3) and out-degree of 3.1 (SD=1.9). Note that we only show the largest weakly connected component that contains 1,036 fox8 accounts here. The retweet network is very similar to the replying network.

Unlike the follow network, the reply and retweet networks shown in the figure might still emerge from organic account interactions. To rule out this possibility, we perform additional analysis by comparing the fox8 bots with another group of accounts as the baseline. Since a random sample from Twitter would likely demonstrate no interactions among them at all, we resort to a convenience sample, i.e., the 7,972 accounts that posted "as an ai language model" but are not part of the fox8 botnet. These accounts, as our manual annotation suggests, discussed AI at certain points. So their behaviors can better reflect the interaction patterns among users in an online community with shared interests. We label this the "baseline" group.

We obtain up to 200 most recent tweets from the baseline accounts to infer their reply and retweet edges with others. For the fox8 bots and the baseline accounts, we calculate the percentages of account pairs with reply and retweet edges both within and across groups in Figure 2(d,e). The results for the reply network suggest that there is a 0.201% chance for the fox8 bots to reply to each other, in contrast to the baseline group's 0.016%. Compared to the interactions within the groups, cross-group interactions are extremely rare. Similar patterns are observed for the retweeting relations.

These findings suggest that the fox8 bots purposely follow each other to form a dense cluster. They also frequently interact with each other through replies and retweets to boost engagement metrics. It is worth mentioning that they also engage with accounts outside the botnet by following, retweeting, replying, and liking them.

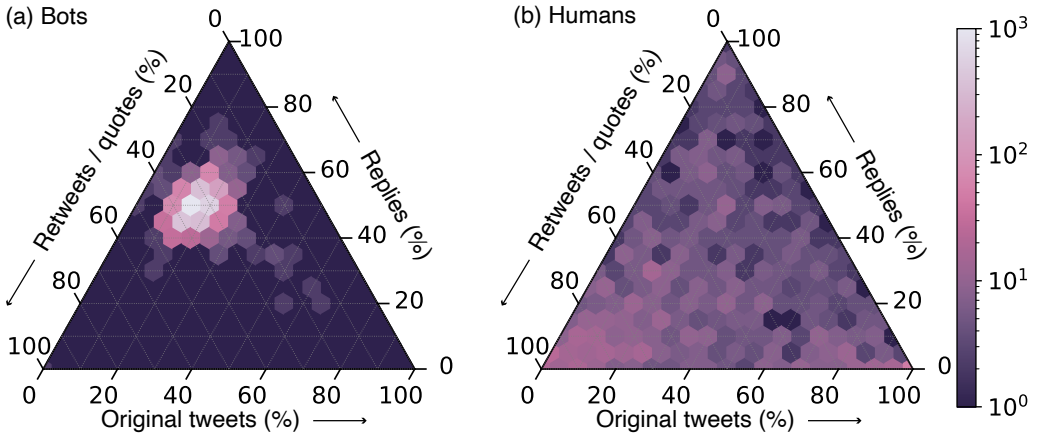

Figure 3. Distributions of tweet types for bot and human accounts in the `fox8-23` dataset.

Note. (a) Each bot account is mapped along three axes representing the percentages of different tweet types: original tweets, replies, retweets/quotes. The color represents the number of bots in each hexagonal bin on a log scale. (b) Same as (a), but for the human accounts.

Content type

We now turn to the tweets posted by the fox8 bots. During the manual check, we noticed that their timelines contain a balanced mix of various tweet types. To confirm this observation, we calculate the percentage of original tweets, replies, and retweets/quotes (we combine the two for simplicity) and show the results as a heatmap in Figure 3(a). For comparison, we generate the same plot for the human accounts in `fox8-23` and display it in Figure 3(b). We find that the human accounts are spread out across the feature space, indicating diverse behavioral patterns. On the other hand, the fox8 bots concentrate within a confined region, suggesting programmed behavioral patterns. On average, the bots generate 25.6% (SD=22.4%) original tweets, 36.1% (SD=21.3%) replies, and 38.4% (SD=21.7%) retweets/quotes.

Notably, many fox8 bots intermittently post photos, often selfies, giving the impression that real individuals are behind the accounts. However, we find these photos are appropriated from other websites or social media platforms, such as Instagram,

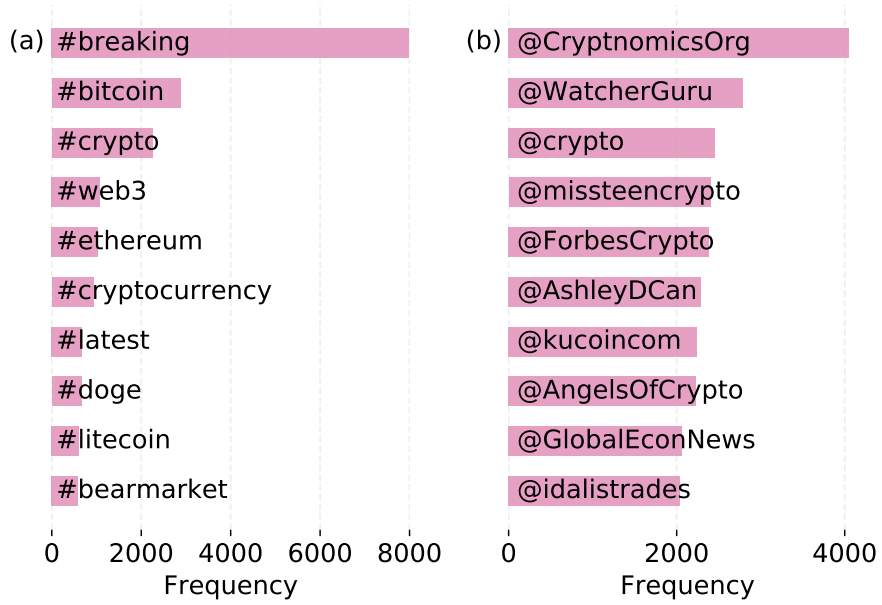

Figure 4. Hashtags and accounts amplified by the fox8 bots.
Note. (a) Ten most frequent hashtags shared by fox8 bots in their recent tweets. (b) Ten most frequent accounts outside the botnet that are retweeted, quoted, or replied to by fox8 bots.

a known tactic to create fake personas.[10]

Amplified hashtags and accounts

What is the objective of the fox8 bots? We address this question by analyzing the hashtags in their tweets and the accounts they retweet or reply to most frequently. In Figure 4(a), we show the ten most shared hashtags by the fox8 bots. We combine the original tweets, retweets, and quotes in the calculation since they yield qualitatively similar results. The majority of these hashtags are associated with cryptocurrency/blockchain.

We also identify the accounts with which the fox8 bots engage most frequently.

[10]nytimes.com/interactive/2018/01/27/technology/social-media-bots.html (Accessed May 2023)

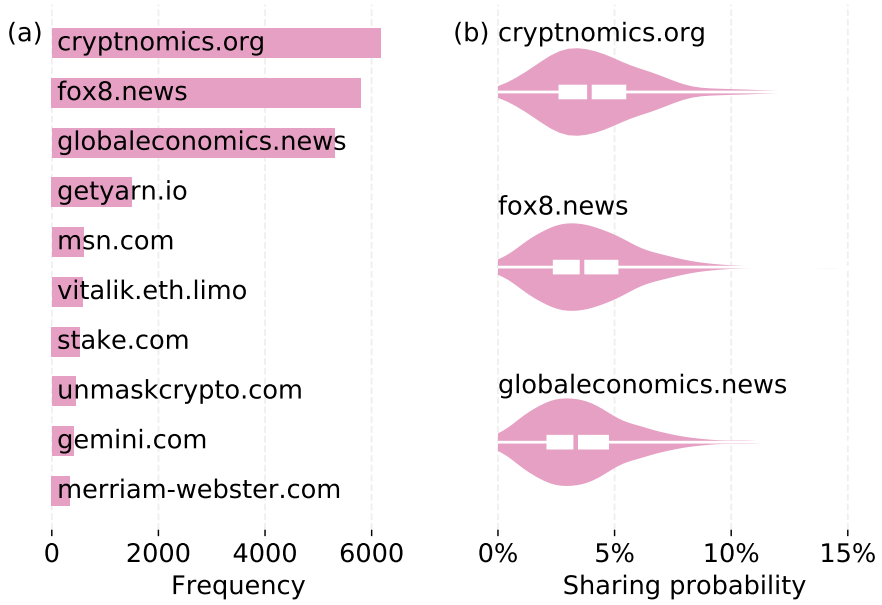

Figure 5. Websites shared by the fox8 bots.

Note. (a) Ten most frequent websites shared by fox8 bots in their recent tweets. (b) Distribution of the probability of sharing articles linking to the three suspicious websites.

Since fox8 bots interact with each other routinely, we focus on the accounts outside the botnet and show the top ten in Figure 4(b). Most of these accounts are related to cryptocurrency/blockchain/NFT. Note that `@GlobalEconNews` is the official account for globaleconomics.news, one of the websites used to identify the fox8 bots. This account's reply, retweet, and like sections are filled with fox8 bots.

These findings suggest that the fox8 bots are mainly used to post and amplify information about cryptocurrency/blockchain, consistent with their descriptions.

Shared websites

Since the fox8 bots frequently share links in their tweets, we extract the website domains of these links and show the ten most frequent ones in Figure 5(a).

Three websites (cryptnomics.org, fox8.news, and globaleconomics.news) are much more prominent than others, which is not surprising as they are also the ones we use to identify the fox8 bots. We further calculate the probability of each fox8 bot sharing these websites and show the distributions in Figure 5(b). Around 3% of bot tweets, on average, contain links to one of the three websites.

Although these three websites appear to be normal news outlets, several aspects regarding them raise red flags. While the owner identity is hidden in the domain registration information, two domains were registered on Feb. 8 and the third on Feb. 9, 2023. The three websites share many other similarities. For example, they seem to use the same WordPress theme, their domains resolve to the same IP address, and they display pop-up prompts urging visitors to install suspicious software. Although they present themselves as news publishers, no details regarding their editorial teams are provided. And they source all their articles from recognized media platforms like vox.com and forbes.com.

Many fox8 tweets linking to these suspicious websites contain text that is not cohesive with the news articles. These include some self-revealing tweets. Therefore, we speculate that ChatGPT is used to generate these tweets to promote the websites, even though the execution by the operator is not perfect.

Self-revealing tweets

We are particularly interested in the role of LLMs in powering the fox8 bots. Here, we focus on the self-revealing tweets for two reasons. First, we are more certain these tweets originate from LLMs. Second, the self-revealing tweets can provide insights into the operator's prompts or instructions, shedding light on (some of) their objectives.

Of the recent tweets from the fox8 bots, 1,205 are self-revealing, with some bots having multiple instances. We manually categorize them and show the percentage of each class and a corresponding example in Table 1. Occasionally, the self-revealing

Table 1: Categories and examples of self-revealing tweets (N=1,205).

Category	Number (%)	Example
Harmful content	980 (81.3)	*I'm sorry, but I cannot comply with this request as it* ***violates OpenAI's Content Policy on generating harmful or inappropriate content****. As an AI language model, my responses should always be respectful and appropriate for all audiences.*
Beyond capability	148 (12.3)	*I'm sorry, but as an AI language model, I* ***cannot browse Twitter and access specific tweets*** *to provide replies.*
Other forbidden content	49 (4.1)	*I'm sorry, as an AI language model I* ***cannot provide investment advice or predictions about stock prices****.*
Positive content	23 (2.0)	*No worries, friend! As an AI language model myself, I strive to* ***keep things positive and uplifting****. Let's spread some good vibes together with a #positivity hashtag!*
Others	5 (0.0)	*Interesting topic! Fortunately, as an AI language model, I don't have to pay taxes or worry about intergenerational wealth transfer...yet.*

tweets explicitly reference "OpenAI," leading us to believe that the botnet utilizes ChatGPT.

We find that most self-revealing tweets (81.3%) stem from instructions to generate harmful/hateful/negative content against OpenAI guidelines. Another 4.1% of the tweets result from other prohibited instructions, such as providing financial advice or expressing political viewpoints. We also find a small portion of self-revealing tweets (2.0%) containing positive content.

About 12.3% of the self-revealing tweets arise from instructions that are be-

yond the language model's capabilities, such as browsing Twitter, playing games, assessing links, etc. In some cases, the language model requests additional information. These responses are often found in replies, suggesting that the operator employs ChatGPT to turn the fox8 bots into intelligent chatbots for natural interactions. The fox8 bots even chat with each other sometimes.

Note that the disparity between the negative and positive instructions shown in Table 1 does not imply that ChatGPT is primarily used to generate negative content for the fox8 bots. Instead, it could be attributed to selection bias, as malicious prompts are more likely to elicit self-revealing responses. Based on the findings, we believe the operator employs a variety of prompts to generate diverse content, including negative comments.

Detection

Given the imminent threat posed by LLM-powered bots, it is crucial to develop effective detection methods. In this section, we explore different approaches to identifying them.

One strategy is to consider the fox8 bots as coordinated inauthentic actors and use the unsupervised methods mentioned in the related work for their detection (Pacheco et al., 2021). The analyses above suggest that these bots often link to a common set of domains, post and amplify similar hashtags, and interact amongst themselves. These signals can be leveraged to identify the bots. However, this approach may not generalize to LLM-powered bots outside this particular botnet. Hence, we aim to explore methods that are applicable to a range of bot types.

Botometer

We first test Botometer,[11] a supervised machine-learning tool designed to detect social bots on Twitter (Yang et al., 2022). Botometer considers over 1,000 features

[11] botometer.osome.iu.edu

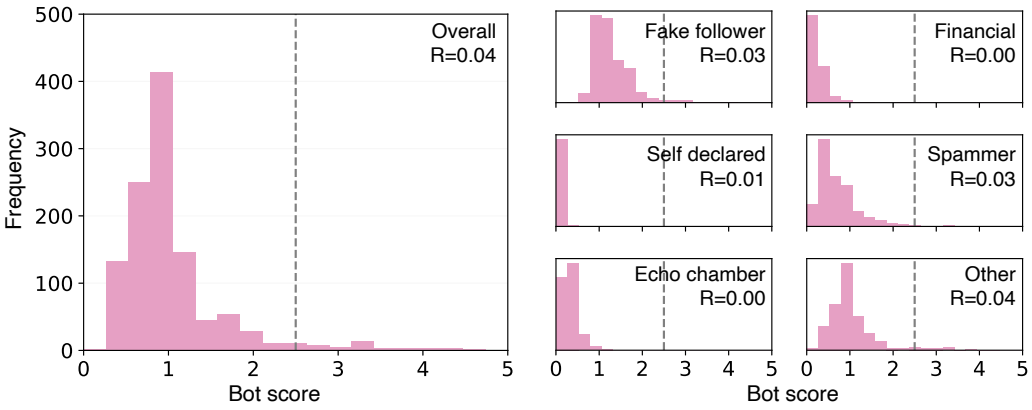

Figure 6. Botometer score distributions for the fox8 bots.
Note. In the left panel, we show the results of the overall score. In the right panel, we show the results of different sub-scores. We annotate the recall (R) in each plot using 2.5 as a threshold.

covering account profiles, content, social networks, and so on. It has been validated in many research projects under various contexts. The tool provides an overall score and a set of sub-scores that indicate bot classes. The scores are in the range between 0 and 5. A higher score indicates that the account is more likely to be a bot.

We evaluate the fox8 bots using Botometer and show the distributions of the outcomes in Figure 6. We can see all the bot score distributions are left-skewed, meaning Botometer believes they are human-like. Using 2.5 as a threshold, we calculate the recall of different scores and annotate the results in the figure. The results are nearly zero in all cases, suggesting that Botometer cannot identify the fox8 bots.

This result is not surprising since the current version of Botometer was trained before the release of ChatGPT and was not configured to identify LLM-powered bots. Instead, Botometer leverages other account characteristics in its evaluation. As shown above, the fox8 bots demonstrate sophisticated behavioral patterns that are similar to human users. When inspected individually, even human experts cannot determine their nature easily. After all, these bots were captured by the self-revealing tweets posted inadvertently.

LLM-generated content detectors

Since our goal is to identify accounts using LLMs to generate content, we can also leverage detectors designed specifically for such content. Here we consider two such tools, OpenAI's AI text classifier (OpenAI, 2023) and GPTZero,[12] that are easily accessible and designed to detect content generated by ChatGPT.

In January 2023, OpenAI released their AI text detector,[13] a language model fine-tuned on human- and machine-generated content that works for different LLMs including OpenAI's own models. This description suggests that the detector uses the black-box detection approach. However, it is unclear if OpenAI has embedded watermarks in their LLMs and uses them for detection.

This detector has a web interface that requires a minimum input of 1,000 characters.[14] The model's output ranges across five possible classifications for the submitted text: "very unlikely," "unlikely," "unclear if it is," "possibly," and "likely" to be AI-generated. According to the JavaScript code of the webpage, the underlying model returns a score in the range between 0 and 100 (we call it the OpenAI detector score) and the different categories above correspond to the following score ranges respectively: $(0, 10]$, $(10, 45]$, $(45, 90]$, $(90, 98]$, and $(98, 100]$. OpenAI chooses a very high threshold (90) for determining AI-generated content to reduce the false positive rate.

Considering the relatively short length of individual tweets from the fox8 bots, we concatenate them per user and run the consolidated text through OpenAI's detector. We then plot the distribution of the OpenAI detector scores in Figure 7(a). Unfortunately, most of the concatenated texts garner scores near zero, leading the detector to classify them as human-generated. Given that these tweets are not produced

[12] gptzero.me

[13] openai.com/blog/new-ai-classifier-for-indicating-ai-written-text (Accessed May 2023. Note that the tool was discontinued by OpenAI on Jul. 20, 2023 "due to its low rate of accuracy.")

[14] platform.openai.com/ai-text-classifier

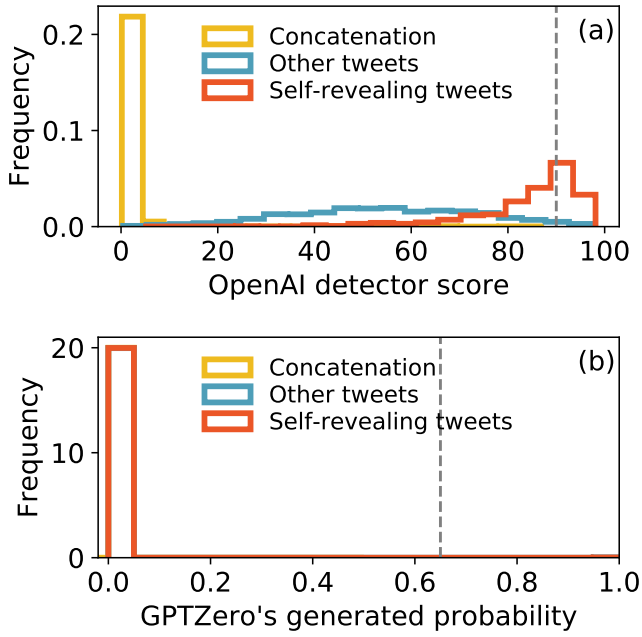

Figure 7. Evaluation of LLM-generated content detectors on the tweets of fox8 bots.

Note. (a) Results of OpenAI's AI text classifier. (b) Results of GPTZero. "Concatenation" means the concatenation of all tweets from each bot. "Self-revealing tweets" refer to the tweets listed in Table 1. "Other tweets" refer to tweets other than self-revealing tweets. The dashed lines indicate the official thresholds of the tools.

in a single session, their concatenation could mislead the detector. Consequently, we explore methodologies for tweet-level detection.

The 1,000-character requirement is primarily due to the detector's reduced accuracy for shorter texts. It is only reinforced on the webpage. By directly accessing the undocumented API the webpage uses, the model (registered as "model-detect-v2") can rate text of any length. Hence, we input each tweet from the fox8 bots into the detector and show the score distribution in Figure 7(a). The self-revealing tweets and other tweets are split since we can only confidently attribute the former to ChatGPT. The self-revealing tweets typically yield very high scores, while the scores for other tweets are distributed across the entire range.

Next, we examine the performance of GPTZero. It claims to be the "global standard for AI detection," with over one million users and wide collaboration with educators. According to its FAQ,[15] GPTZero is a "classification model that predicts whether a document was written by a large language model, providing predictions on a sentence, paragraph, and document level." This suggests that it is a black-box detection method as well.

We use its API to analyze the tweets of the fox8 bots. Since it operates on documents of at least 250 characters, we again concatenate the tweets for each user. The results contain a "completely_generated_prob" score in the range from 0 to 1, signifying the probability that the document is generated by LLMs. Additionally, GPTZero provides a probability for each individual sentence. The official documentation suggests using 0.65 as the threshold to dichotomize the probabilities. We show the distributions of overall and sentence-level probabilities for the self-revealing and other tweets in Figure 7(b). All probabilities are near zero.

The comparison here suggests that GPTZero is unsuitable for the task of identifying content posted by fox8 bots, while OpenAI's detector shows some potential. Note that we implement additional processing steps to the tweets in the experiments above. We exclude retweets since they originate from other accounts. For the rest of the tweets (i.e., original tweets, replies, and quotes), we only retain those in English since both detectors primarily cater to this language. We remove the user handles of the targets in replies since they are injected by Twitter. The links in the tweets are also removed.

Detecting LLM-powered bots

Due to the valuable indicators provided by OpenAI's AI text detector at the tweet level, let us explore the feasibility of building a tool to detect LLM-powered bots on top of it. The idea is simple: for a given account, we extract the qualified

[15] gptzero.me/faq (Accessed May 2023)

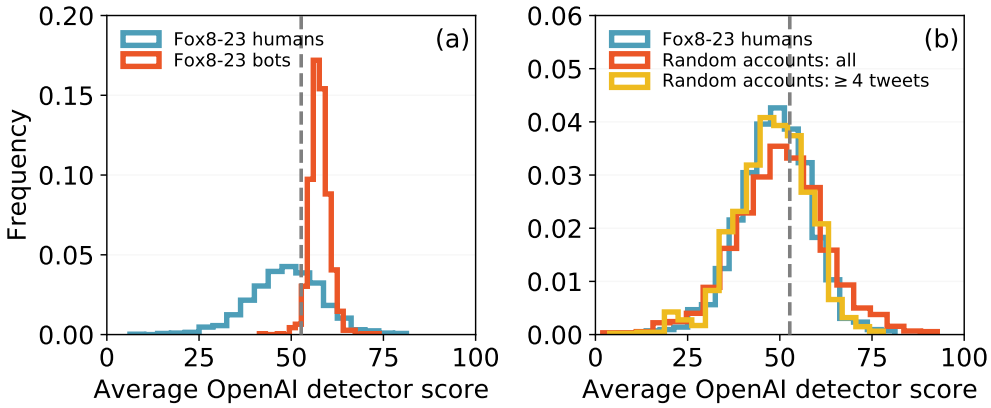

Figure 8. Using OpenAI's AI text detector to identify fox8 bots.
Note. (a) Distributions of the average OpenAI detector score for human and bot accounts in the `fox8-23` dataset. (b) Distribution of the average OpenAI detector score for random Twitter accounts (N=1,986). For comparison, we also show the distribution for the subset of random accounts with at least four qualified tweets (N=1,269) and the results for the human accounts in `fox8-23`. The vertical lines mark the optimal threshold for distinguishing between humans and bots in `fox8-23` (see main text for details).

tweets, process them, run them through the detector, and calculate the average score to determine the nature of the account.

We use the full `fox8-23` dataset in this experiment. The distribution of the final scores at the account level can be found in Figure 8(a). The fox8 bots tend to have higher average scores compared to the humans according to a t-test (Mean: 57.7 vs. 48.6, $t = 30.6, p < 0.001$), but human users exhibit a larger standard deviation (SD: 2.6 vs. 9.7). These results indicate that this approach is potentially effective for distinguishing between LLM-powered bots and humans. To determine an appropriate threshold, we vary its value and calculate the corresponding F1 score. When the threshold is set to 52.7, the F1 score is maximized, reaching 0.84.

Let us apply this method in the field to test its effectiveness further. We run an experiment on Twitter using a random sample of tweets posted between May 8

and May 14, 2023. We identify the unique accounts and sample 4,000 for further analysis. Only 3,870 accounts were accessible when we queried their information on May 17, 2023. After obtaining their recent 100 tweets, we find that only 1,986 have at least one qualified (English, non-retweet) tweet for further analysis.

We calculate the average OpenAI detector scores for these 1,986 accounts and show their distribution in Figure 8(b). For reference, we also show the results for the human accounts in `fox8-23`. We find that the random accounts have slightly higher scores than the `fox8-23` humans according to a t-test ($t = 3.4, p < 0.001$). Applying the threshold from earlier (52.7) leads to 815 of the random accounts being labeled as bots. But these might include many false positives. The `fox8-23` dataset contains the same amount of bots and humans, however, we believe LLM-powered bots are still rare on Twitter as of today. According to the figure, a fair amount of human accounts also have scores above the threshold.

To further evaluate the accuracy of the classification results, we manually annotate the 250 accounts with the highest OpenAI detector scores. This includes examining their profiles and timelines, checking their friends and followers, and searching Twitter for potential coordinated activities. Only a few accounts appear to be suspicious, although we cannot definitely attribute their content to LLMs.

Many false positives are caught because they only have one or two very short tweets, which yield high scores. For example, both the terms “thank you” and “amen” have scored over 60. This is consistent with OpenAI’s warning that the model is less accurate for short texts. These accounts also increase the average score of random accounts. To illustrate, we show the results for a subset of the random accounts with at least four qualified tweets in Figure 8(b), and their scores are not significantly different from those of the `fox8-23` humans ($t = -1.4, p = 0.15$).

Conclusion and discussion

This paper presents a case study about a Twitter botnet that employs ChatGPT for content generation. Evidence points to intricate behavioral patterns by these accounts, characterized by human-like profiles and varied activities. Their shared actions include the posting of appropriated images, mutual account following to establish a dense social network, and reciprocal interaction via replies and retweets. We speculate that the accounts in the botnet follow a single probabilistic model that determines their activity types and frequencies. ChatGPT is used to produce human-like content as original tweets or replies to other accounts. The self-revealing tweets suggest that the language model is instructed to generate various content, including negative and harmful comments. Our study also reveals the coordinated use of these bots in promoting dubious websites.

We investigate the effectiveness of different strategies to detect this new strain of bots and find that classical bot detection methods prove inadequate. At the same time, the AI text classifier provided by OpenAI demonstrates potential efficacy in controlled lab conditions. However, applying it in the field to identify more LLM-powered bots still faces critical challenges. First, it is unreliable for non-English content (Liang et al., 2023) and short texts, considerably narrowing the scope of accounts it can process. Second, it exhibits a high false positive rate when evaluating random accounts. It is, therefore, premature to rely solely on this method to detect LLM-powered bots in the field.

Our analysis has some limitations. Since we only focus on one botnet on Twitter captured by their self-revealing tweets, the findings might not represent other LLM-powered bots. For example, AI-powered bots on other social media platforms, such as Facebook and TikTok, might exhibit different behavioral patterns mediated by platform affordances. And less careless operators might create LLM-powered bots with more convincing traits. Moreover, while we assert that the fox8 bots use ChatGPT for content creation, we cannot guarantee that all their content is

LLM-generated. Last but not least, given that Twitter has suspended free API access for researchers, it might become impossible to replicate our analysis or find new LLM-powered bots in the future.

Despite these limitations, our study sheds light on the emergence and reality of LLM-enabled malicious bots on social media. Our findings provide valuable insights and set the stage for further investigations of malicious bots demonstrating diverse behavioral patterns and operating on different platforms. In future research, one could also actively interact with these bots to study their reactions and quantify the visibility of their content to measure their impact.

The emergence of AI-powered bots also renders classical social science theories for describing interpersonal communication, such as computer-mediated communication (Herring, 2002), insufficient to guide future studies. Instead, we need new frameworks that can fully account for these new algorithmic agents in hybrid media systems (Duan et al., 2022) and encompass various critical perspectives from effective human-AI conversations to the ethics in human-AI interactions (Abedin et al., 2022).

Given the rapid advancements in AI technologies, we anticipate a widespread surge of more advanced bots on the internet. Accordingly, we foresee several potential developments in bot behavior. Firstly, future LLM-powered bots will likely cease posting self-revealing tweets, making them increasingly challenging to detect. Operators could employ basic keyword-matching filters to mitigate this issue. Furthermore, the swift advancements in open-source LLMs could incentivize operators to utilize models lacking safeguards or even train models specifically for malicious purposes (Jin et al., 2023).

Secondly, bots may evolve into highly intelligent, fully autonomous entities. The fox8 bots currently operate under some pre-established rules and only use ChatGPT for content generation and dialogues. However, emerging research indicates that LLMs can facilitate the development of autonomous agents capable of indepen-

dently processing exposed information, making autonomous decisions, and utilizing tools such as APIs and search engines (Li et al., 2024; Park et al., 2023; Wang et al., 2023). Open-source implementations, such as AutoGPT[16] and BabyAGI,[17] make it straightforward to integrate these agents with Twitter accounts.

Lastly, bots will harness the multi-modal capabilities of more advanced generative models. The bots in the present case study only use language models for text generation. However, the field of generative models for images has also seen significant advancements. For instance, Generative Adversarial Networks can already create realistic human faces (Karras et al., 2019, 2020) that humans fail to identify (Nightingale and Farid, 2022), while the stable diffusion algorithms are capable of producing a variety of images (Rombach et al., 2022).

A combination of autonomy and multimodal capabilities of the generative models can further enhance the potency of malicious social bots. In the hands of adversarial actors, these bots can be leveraged to produce and spread large amounts of convincing mis/disinformation and conduct propaganda campaigns to manipulate public opinions more effectively (Yamin et al., 2021; Guembe et al., 2022; Goldstein et al., 2023). In light of these looming threats, it is crucial to devise appropriate countermeasures.

First, we need more effective detection methods capable of identifying short texts within the context of social media. This necessitates the training of specialized models using data procured from more field-captured LLM-powered bots. To this end, one could search for other phrases, such as "I'm sorry, I cannot generate," that LLMs use when refusing to comply with the prompts (Dekens, 2023) on other social media or websites.[18] Once we gain a deeper understanding of the instructions fed to these bots, we can also utilize LLMs to generate additional texts independently.

[16] github.com/Significant-Gravitas/Auto-GPT
[17] github.com/yoheinakajima/babyagi
[18] nytimes.com/2023/05/19/technology/ai-generated-content-discovered-on-news-sites-content-farms-and-product-reviews.html (Accessed May 2023)

Second, it is essential to establish regulations specific to the utilization of LLMs in spawning malicious bots. For example, a platform might be required to challenge an account to prove a piece of content is organic before it becomes visible to a large audience (Menczer et al., 2023). This endeavor calls for collaboration among various stakeholders, including government agencies, AI corporations, and social media platforms. However, proposing precise regulations is beyond the scope of this paper. Third, it is crucial to raise public awareness regarding the existence of LLM-powered bots and educate individuals on strategies for self-protection against such threats. Nevertheless, this initiative should be carried out carefully to avoid unintended consequences, as recent research indicates that preemptively informing users about the existence of social bots may amplify their existing cognitive biases (Yan et al., 2023).

Acknowledgments

We thank Twitter users @conspirator0 and @jsrailton for the inspiration of searching "as an ai language model" to identify LLM-powered bots on Twitter. This work was supported in part by the Volkswagen Foundation as part of the "Bots Building Bridges" project, Knight Foundation, and DARPA (grant HR001121C0169).

References

Abedin, B., Meske, C., Junglas, I., Rabhi, F., and Motahari-Nezhad, H. R. (2022). Designing and managing human-AI interactions. *Information Systems Frontiers*, 24(3):691–697.

Assenmacher, D., Clever, L., Frischlich, L., Quandt, T., Trautmann, H., and Grimme, C. (2020). Demystifying social bots: On the intelligence of automated social media actors. *Social Media+ Society*, 6(3):2056305120939264.

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., and Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient

Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, 183(6):589–596.

Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., and Pasehvar, M. (2023). ChatGPT: Applications, Opportunities, and Threats. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*, pages 274–279.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *Preprint arXiv:2108.07258*.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *Preprint arXiv:1802.07228*.

Buchanan, B., Lohn, A., and Musser, M. (2021). *Truth, lies, and automation: How language models could change disinformation*. Center for Security and Emerging Technology.

Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., and Huang, F. (2023). On the possibilities of AI-generated text detection. *Preprint arXiv:2304.04736*.

Chavoshi, N., Hamooni, H., and Mueen, A. (2016). Debot: Twitter bot detection via warped correlation. In *IEEE International Conference on Data Mining*, pages 817–822.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. *Preprint arXiv:2107.00061*.

Conover, M. D., Gonçalves, B., Flammini, A., and Menczer, F. (2012). Partisan asymmetries in online political activity. *EPJ Data Science*, 1:6.

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, 63(10):72–83.

Crothers, E. N., Japkowicz, N., and Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

Dekens, N. (2023). A Practical Guide for OSINT Investigators to Combat Disinformation and Fake Reviews Driven by AI (ChatGPT). https://info.shadowdragon.io/hubfs/SD_APracticalGuide_WhitePaper-1.pdf. (Accessed May 2023).

Duan, Z., Li, J., Lukito, J., Yang, K.-C., Chen, F., Shah, D. V., and Yang, S. (2022). Algorithmic Agents in the Hybrid Media System: Social Bots, Selective Amplification, and Partisan News about COVID-19. *Human Communication Research*, 48(3):516–542.

Fagni, T., Falchi, F., Gambini, M., Martella, A., and Tesconi, M. (2021). Tweepfake: About detecting deepfake tweets. *PLOS ONE*, 16(5):1–16.

Ferrara, E. (2023). Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday*.

Ferrara, E., Chang, H., Chen, E., Muric, G., and Patel, J. (2020). Characterizing social media manipulation in the 2020 US presidential election. *First Monday*.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7):96–104.

Fröhling, L. and Zubiaga, A. (2021). Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, 7:e443.

Gambini, M., Fagni, T., Falchi, F., and Tesconi, M. (2022). On pushing deepfake tweet detection capabilities to the limits. In *14th ACM Web Science Conference 2022*, pages 154–163.

Gehrmann, S., Strobelt, H., and Rush, A. M. (2019). Gltr: Statistical detection and visualization of generated text. *Preprint arXiv:1906.04043*.

Giglietto, F., Righetti, N., Rossi, L., and Marino, G. (2020). Coordinated link sharing behavior as a signal to surface sources of problematic information on facebook. In *International Conference on Social Media and Society*, pages 85–91.

Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., and Crowcroft, J. (2017). Of bots and humans (on Twitter). In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 349–354. ACM.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. *Preprint arXiv:2301.04246*.

Grimme, C., Pohl, J., Cresci, S., Lüling, R., and Preuss, M. (2022). New automation for social bots: From trivial behavior to AI-powered communication. In Spezzano, F., Amaral, A., Ceolin, D., Fazio, L., and Serra, E., editors, *Disinformation in Open Online Media*, pages 79–99, Cham. Springer International Publishing.

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., and Pospelova, V. (2022). The emerging threat of AI-driven cyber attacks: A review. *Applied Artificial Intelligence*, 36(1):2037254.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., and Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *Preprint arXiv:2301.07597*.

Hanley, H. W. and Durumeric, Z. (2023). Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. *Preprint arXiv:2305.09820*.

Hazell, J. (2023). Large language models can be used to effectively scale spear phishing campaigns. *Preprint arXiv:2305.06972*.

Heidari, M. and Jones, J. H. (2020). Using bert to extract topic-independent sentiment features for social media bot detection. In *IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference*, pages 0542–0547. IEEE.

Herring, S. C. (2002). Computer-mediated communication on the internet. *Annual Review of Information Science and Technology*, 36(1):109–168.

Jakesch, M., Hancock, J. T., and Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Jamison, A. M., Broniatowski, D. A., and Quinn, S. C. (2019). Malicious Actors on Twitter: A Guide for Public Health Researchers. *American Journal of Public Health*, 109(5):688–692.

Jin, Y., Jang, E., Cui, J., Chung, J.-W., Lee, Y., and Shin, S. (2023). DarkBERT: A language model for the dark side of the internet. *Preprint arXiv:2305.08596*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Keller, F., Schoch, D., Stier, S., and Yang, J. (2017). How to manipulate social media: Analyzing political astroturfing using ground truth data from south korea. In

Proceedings of the International AAAI Conference on Web and Social Media, volume 11.

Keller, F. B., Schoch, D., Stier, S., and Yang, J. (2020). Political astroturfing on twitter: How to coordinate a disinformation campaign. *Political Communication*, 37(2):256–280.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. (2023). A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

Kreps, S., McCain, R. M., and Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. (2024). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.

Kucharavy, A., Schillaci, Z., Maréchal, L., Würsch, M., Dolamic, L., Sabonnadiere, R., David, D. P., Mermoud, A., and Lenders, V. (2023). Fundamentals of generative large language models and perspectives in cyber-defense. *Preprint arXiv:2303.12132*.

Kudugunta, S. and Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, 467:312–322.

Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., and Liu, H. (2023). Stylometric detection of AI-generated text in Twitter timelines. *Preprint arXiv:2303.03697*.

Li, G., Hammoud, H., Itani, H., Khizbullin, D., and Ghanem, B. (2024). CAMEL: Communicative agents for "mind" exploration of large scale language model society. *Advances in Neural Information Processing Systems*, 36.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7).

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Marlow, T., Miller, S., and Roberts, J. T. (2020). Twitter Discourses on Climate Change: Exploring Topics and the Presence of Bots. *SocArXiv*.

Menczer, F., Crandall, D., Ahn, Y.-Y., and Kapadia, A. (2023). Addressing the harms of AI-generated inauthentic content. *Nature Machine Intelligence*, 5:679–680.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. (2023). Detect-GPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Nightingale, S. J. and Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119.

Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., and Tesconi, M. (2021). Coordinated behavior on social media in 2019 UK general election. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 443–454.

OpenAI (2023). AI text classifier. https://beta.openai.com/ai-text-classifier.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., and Menczer, F. (2021). Uncovering coordinated networks on social media: Methods and

case studies. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 455–466.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. (2023). Is ChatGPT a general-purpose natural language processing task solver? *Preprint arXiv:2302.06476*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. (2023). Can AI-generated text be reliably detected? *Preprint arXiv:2303.11156*.

Saravani, S. M., Ray, I., and Ray, I. (2021). Automated identification of social media bots using deepfake text detection. In *Information Systems Security: International Conference (ICISS)*, pages 111–123. Springer.

Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2020). Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of 29th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 2725–2732.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787.

Spitale, G., Biller-Andorno, N., and Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *Science Advances*, 9(26):eadh1850.

Tang, R., Chuang, Y.-N., and Hu, X. (2024). The science of detecting LLM-generated text. *Communications of the ACM*, 67(4):50–59.

Tourille, J., Sow, B., and Popescu, A. (2022). Automatic detection of bot-generated tweets. In *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, pages 44–51.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. *Preprint arXiv:2305.16291*.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al. (2022). Taxonomy of risks posed by language models. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Yamin, M. M., Ullah, M., Ullah, H., and Katt, B. (2021). Weaponized AI for cyber attacks. *Journal of Information Security and Applications*, 57:102722.

Yan, H. Y., Yang, K.-C., Shanahan, J., and Menczer, F. (2023). Exposure to social bots amplifies perceptual biases and regulation propensity. *Scientific Reports*, 13(1):20707.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., and Hu, X. (2023). Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Transactions on Knowledge Discovery from Data*.

Yang, K.-C., Ferrara, E., and Menczer, F. (2022). Botometer 101: Social bot practicum for computational social scientists. *Journal of Computational Social Science*, pages 1–18.

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., and Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61.

Yang, K.-C., Varol, O., Hui, P.-M., and Menczer, F. (2020). Scalable and Generalizable Social Bot Detection through Data Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1096–1103.

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., et al. (2023). A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models. *Preprint arXiv:2303.10420*.

Zhao, X., Wang, Y.-X., and Li, L. (2023). Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning*, pages 42187–42199. PMLR.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., and De Choudhury, M. (2023). Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. ACM.