CS492 Winter 2024
Final Project Milestone #2 - Progress Report
March 15, 2024
Alex Hejmej [20779600], Tailai Wang [20816387], Isaiah Witzke [20845973]
Link to GitHub Repository

---

# Interactive Storytelling Website about AI Bot Toxicity on Social Media Platforms

## Summary: What are the goals of the project?

The driving theme of our project remains the same as Milestone 1 - we are curious about how bots cause harm on social media platforms, and we're eager to discover how to effectively manage them and their impact. The 3 Phase Approach we proposed still applies:

1. Analysis & Synthesis of Existing Literature / Content
    a. We've completed research on most of the topics of interest we mentioned in our proposal, and have compiled some key insights below in Section 4.
    b. We started documenting our readings and findings in our GitHub repository (linked above)
2. Applying Research Findings in a Data-Scraping Investigation
    a. Through our Phase 1 research, we've been able to more specifically define the scope of this Data-Scraping Investigation. We're going to focus on scraping posts and accounts aggregated using the Python Reddit API Wrapper, and then using open-source bot-detection tools on the data. We cover the inspiration for this decision in Sections 2 and 4.
    b. We started breaking down the tasks required to do this Data-Scraping Investigation as Issues in our GitHub Repository (linked above).
3. Display Research Synthesis and Experiment Results on an Interactive Website
    a. We will begin this last phase of the project in the weeks after Milestone 2 is handed in. The Issues we compiled on our GitHub repository planned out our process for designing the features, writing the copy, and actually developing/deploying the website.
    b. Our immediate next steps are finalizing what technologies and libraries to use and getting started with some boilerplate code.

## Exploration: What questions are the team trying to answer?

In our Milestone 1 Project Proposal, we outlined the research questions we wanted to explore in Phase 1. These are covered in-depth in Section 4. As we completed this research, we realized that there are some new additional questions that we'd like to explore in Phase 2 of our project:

a. Why doesn't Reddit already have automated bot detection and better moderation tools like other social media platforms?
    i. Our initial research in Section 4 Part d showed us how lackluster Reddit's efforts are compared to sophisticated trust and safety orgs like Meta and Twitter. As the company approaches its long-awaited Initial Public Offering, we'd like to conduct some due diligence into why its tooling is so behind other platforms.
    ii. Colloquially, the job of "Subreddit moderator" involves very thankless and mundane work. We're interested in improving that experience.
b. What options are available for Reddit to create better tools for its moderators?

   i. Our Data-Scraping investigation will enable us to test and compare some open-source bot detection tools.
   ii. We'll present the results of this investigation on our website as a proposal for Reddit to offer better tools for its moderators.

<u>Justification: Why is the form of the project appropriate to answer those questions?</u>

Given the data-intensive nature of our investigation, the most direct and accessible way to display our findings is through an interactive website. On top of providing information, we think the website can be seen as a persuasive tool to pitch our findings to Reddit. If a figurative Reddit employee saw how effective certain tools were at uncovering bots, they might be inspired to implement our research methodology in production.

<u>Evidence: What research has the team undertaken thus far?</u>

In our Milestone 1 Project Proposal, we outlined a few areas of research to complete in Phase 1 of our project. Below are some of the summarized findings from that research:

1. *Understanding the prevalence of the bot problem*
   a. ABC news quotes that cybersecurity firm CHEQ found that X, formerly known as Twitter, has 75% of its traffic coming from bots, while other platforms like Facebook, Instagram, and TikTok have only around 3% (Purtill, 2024). This aligns with the fact that the owner of Twitter, Elon Musk, cut the moderation costs substantially as well dissolved the Trust and Safety Council when he took over, sparking an influx of bots on the platform (López Restrepo, 2024). Elon's attempts at resolving the bot crisis by introducing a new paid account tier have not been effective either due to the ranking bias incentive that made more bots pay for this tier.
   b. Generative AI such as GPT-3 and GPT-4 has also exacerbated the problem by making these bots more believable to users and harder to filter out by bot detection algorithms (Douglas Heaven, 2020). One of the larger intersecting problems between bots, generative AI, and these social media platforms is the growing portion of "For You" content, algorithms that suggest content to users from people outside their social networks (Hu, 2024). With this new prioritized form of content for social media sites, generative AI bots have much more potential to influence internet users.
   c. It should be noted that one of the challenges in researching the prevalence of bots across platforms is since bot detection is challenging to begin with, it is hard to find a consensus on the exact percentage of users who are bots and those who are real for each platform (Walsh, 2023). The statistic from CHEQ above is just one organization's estimates.
2. *Who is creating these bots and why?*
   a. There are many different bad actors that use these bots, but their motives usually fall into two categories: foreign influence and financial gain.
   b. Over the course of the Ukraine-Russia war, many bot accounts have been found supporting Pro-Russian opinions to improve public perception of Russia's invasion of Ukraine (Press Association, 2022). This botfarm (i.e. a network of bots) has been linked back to an old arms factory in St Petersburg that is suspected to be run by Yevgeny Prigozhin who is accused of meddling in the 2016 US election. These bots not only post their own Pro-Russian content, but amplify genuine support from legitimate accounts as well. Russia is not the only country accused of using these bots to create foreign influence either. China has been accused of targeting Canadian politicians' social media

accounts by leaving thousands of comments in an attempt to silence criticism of Beijing (BBC, 2023).

c. Porn bots have been around for a while, with many users taking notice of them even in 2019 (Tait, 2019). These accounts can make money off of the vast amount of lonely men on the internet in various ways, including leading them to adult websites or scamming these individuals of money in exchange for some hope of intimacy. Some marketing firms even pay other people for leads on these potential customers, paying them once the lead gives them their email. With the rise of deepfakes and generative AI, it's now easier than ever to produce adult content at scale and make a load of money doing so, further incentivizing people to create bots to promote fake accounts with little risk (Hunter, 2024). Other common types of bots try to either steal from you by either getting you to buy crypto coins or NFTs, or try to drive up the price of the coins they themselves own (Uhlemann, 2023). Some bots try to make fake ads look real (which contain malware) by commenting on them with recommendations for the product or solution (Germain, 2023).

3. *Who is being affected by these bots and what are some of the negative outcomes?*
    a. Our research for this part shows that while bots may affect the overall user experience on these platforms and cause frustration, there are much deeper problems (Reddit User quinn_thomas, 2023). Members of other countries such as Russia can utilize these bots to try and undermine democracy by spreading misinformation (Press Association, 2022). With generative AI, this misinformation can be even more powerful and threaten even the belief in objective truths which democracy relies on (Brandt, 2023).
    b. This “liar’s dividend” threatens not only the democracy of countries by making individuals less likely to believe real human-made content, but also threatens the use of our internet altogether (Chesney, 2018). In late February, many people on the internet were believing that a real video showing Black exchange students being discriminated against while fleeing Ukraine was in fact fake, showing that this “liar’s dividend” is already taking effect (Jerkins, 2022).
    c. Of course, users are also at risk of being scammed and having their money, or even identity, taken from them by these bots. The FTC found that in 2022, more than 46,000 people fell to crypto scams, losing over $1 billion in crypto to these scams (Fletcher, 2022). With many bot sighted on twitter promoting their own crypto assets, it seems likely that a significant portion of these scams started on social media.
4. *How do different social media platforms address the bot problem?*
    a. Meta utilizes AI for detecting hate speech as well as identifying bots and spam accounts. Their methods additionally involve human reviews to enhance their AIs’ actions, although specifics are not detailed (Cruz, 2023) (Instagram, 2024).
    b. Reddit relies heavily on volunteer moderators (mods) for community moderation, with limited automated tools for handling AI-generated spam (Clarke, 2023) (Reddit, 2020). Reddit bans are primarily report-based, with fewer automatic bot detection tools compared to other platforms. For example, popular bot-detection projects for Reddit like BotDefense have been shut down after recent changes to Reddit’s API usage limits (Reddit User dequeued, 2023).
    c. Unlike Reddit, Twitter’s APIs have allowed for many public 3rd party bot-detection tools. Algorithms like Botometer and BotOrNot have served as the basis for much literature on

automated bot detection ML algorithms (Ferrara et al., 2017) (Davis et al., 2016) (Kudugunta & Ferrara, 2018), while tools like Bot Sentinel allow users to see misinformation spreads with nefarious accounts (Bouzy, 2022). Although Twitter's in-house approach to detecting and banning harmful bots remains undisclosed, Twitter has criticized third-party bot detection methods as being "extremely limited" (Roth & Pickles, 2020), suggesting that Twitter's bot detection algorithms are more sophisticated than than the methods found in public literature (it should be noted that Twitter's struggle with nefarious bots still persists).

i. 3rd party researchers are able to train bot-detection algorithms reaching 90% accuracy fairly easily. However, as pointed out in recent MIT research, these models will generally perform poorly on unfamiliar "real-world" datasets. One of the primary reasons for this is due to the lack of high-quality datasets on which to train ML algorithms (Walsh, 2023).

5. *What are common industry KPIs for measuring bots and the problem(s) they cause?*

a. Meta is (somewhat surprisingly) the industry leader in publishing comprehensive KPIs for how to measure bots and their efforts against thwarting them. Their Trust and Safety Organization publishes a well-documented transparency website that highlights their efforts in automatically detecting bots and hateful content. Some common KPIs include the percentage of Monthly Active Users (MAU) that is composed of fake accounts, gross volume of fake accounts, and percentage of fake accounts that were automatically removed without user reporting (Meta, 2023).

b. The team can use these KPIs in measuring the tests of the tools in Phase 2 of this project, enabling us to compare our experiment results with the current industry-leading metrics. This should add to the credibility of our project in a recommendation to Reddit.

Reflection: What has the team has learned thus far?

Through Phase 1 of the project, the team has become well-read on the current landscape of bots on social media platforms. We have developed an understanding of impacts and scale of the bot problem, along with the best-in-class solutions to address them. We believe this puts us in a good position for Phase 2 of the project, where we'll apply our knowledge to this Data-Scraping investigation that will result in a tangible recommendation for Reddit.

Reflection: Understanding the connection to course content

A piece of feedback we received in Milestone 1 was to connect the project more with course goals. At a high level, we know that the project is related to the course objective of making informed judgements about the social and ethical consequences of the deployment of computing technologies. More specifically though, we see our project directly supporting the content of class 11: "Fixes: law, regulation, and other approaches". Phase 2 of our project is similar to the two case studies assigned in class 11: the UPenn Paper on letting AI identify problems (Eisenstadt, 2021), and the Algorithmic Justice League blog on bug bounties for algorithmic harms (Kenway, 2022). We believe our recommendation to Reddit after conducting the Data-Scraping Investigation will be inspired by these papers, thus meeting the requirement of connecting the project to course content.

## References

BBC. (2023, October 2023). *Chinese bots targeted Trudeau and others - Canada*. BBC. https://www.bbc.com/news/world-us-canada-67201927

Bouzy, C. (2022). *About*. Bot Sentinel. https://botsentinel.com/info/about

Brandt, Jessica. (2023, November 8). *Propaganda, foreign interference, and generative AI.* Brookings. https://www.brookings.edu/articles/propaganda-foreign-interference-and-generative-ai/

Chesney, Robert and Citron, Danielle Keats (July 14, 2018). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security* . 107 California Law Review 1753 (2019), U of Texas Law, Public Law Research Paper No. 692, U of Maryland Legal Studies Research Paper No. 2018-21, Available at SSRN: https://ssrn.com/abstract=3213954 or http://dx.doi.org/10.2139/ssrn.3213954

Clarke, L. (2023, April 11). Redd*it moderators brace for a ChatGPT spam apocalypse.* Vice. https://www.vice.com/en/article/jg5qy8/reddit-moderators-brace-for-a-chatgpt-spam-apocalypse

Cruz, J. D. (2023, December 14). *Meta launches new tools on Instagram to regulate spam content, bots*. Tech Times. https://www.techtimes.com/articles/299732/20231214/meta-launches-new-tools-instagram-regulate-spam-content-bots.htm

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016, February 2). *Botornot: A system to evaluate social bots*. arXiv.org. https://arxiv.org/abs/1602.00975

Douglas Heaven, William. (2020, October 8). *A GPT-3 bot posted comments on Reddit for a week and no one noticed*. MIT Technology Review. https://www.technologyreview.com/2020/10/08/1009845/a-gpt-3-bot-posted-comments-on-reddit-for-a-week-and-no-one-noticed/

Eisenstadt, Leora. (2021). *#METOOBOTS AND THE AI WORKPLACE.* University of Pennsylvania Journal of Business Law. https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=1660&context=jbl

Ferrara, E., Varol, O., Davis, C. B., Menczer, F., & Flammini, A. (2017, June 19). *The rise of Social Bots*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2982515

Fletcher, Emma. (2022, June 3). *Reports show scammers cashing in on crypto craze*. FTC. https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/06/reports-show-scammers-cashing-crypto-craze

Germain, Thomas. (2023, September 6). W*ant to See How Bad Twitter's Bot Problem Is? Ask for Crypto Help*. Gizmodo. https://gizmodo.com/twitter-bot-problem-metamask-support-crypto-elon-musk-1850808172

Hu, Charlotte. (2024, February 13). *How AI Bots Could Sabotage 2024 Elections around the World*. Scientific American. https://www.scientificamerican.com/article/how-ai-bots-could-sabotage-2024-elections-around-the-world/

Hunter, Tatum. (2024, February 25). *AI 'dream girls' are coming for porn stars' jobs*. The Washington Post. https://www.washingtonpost.com/style/of-interest/2024/02/25/ai-porn-avn-industry/

Instagram. (2024). *How Instagram uses artificial intelligence to moderate content*. Help center. https://help.instagram.com/423837189385631

Jerkins, Morgan. (2022, October). *Black or Bot? The Long, Sordid History of Co-opting Blackness Online*. Mother Jones.

https://www.motherjones.com/media/2022/09/disinformation-russia-trolls-bots-black-culture-blackness-ukraine-twitter/

Kenway, Josh. (2022, January). *Bug Bounties for Algorithmic Harms*?. Algorithmic Justice League. https://www.ajl.org/bugs

Kudugunta, S., & Ferrara, E. (2018, February 18). *Deep Neural Networks for BOT detection*. arXiv.org. https://arxiv.org/abs/1802.04289

López Restrepo, Manuela. (2024, March 4). *How the porn bots took over Twitter*. NPR. https://www.npr.org/2024/03/07/1235784919/twitter-x-bots-social-media-elon-musk

Meta. (2023). *Community Standards Enforcement Report - Fake Accounts*. Meta Transparency Center. https://transparency.fb.com/reports/community-standards-enforcement/fake-accounts/facebook/

Press Association. (2022, May 1). '*Troll factory' spreading Russian pro-war lies online, says UK*. Guardian News and Media. https://www.theguardian.com/world/2022/may/01/troll-factory-spreading-russian-pro-war-lies-online-says-uk

Purtill, James. (2024, February 27). *Twitter is becoming a 'ghost town' of bots as AI-generated spam content floods the internet*. ABC News Australia. https://www.abc.net.au/news/science/2024-02-28/twitter-x-fighting-bot-problem-as-ai-spam-floods-the-internet/103498070

Reddit User dequeued. (2023, July 5). *BotDefense is wrapping up operations*. reddit. https://www.reddit.com/r/BotDefense/comments/14riw76/botdefense_is_wrapping_up_operations

Reddit User quinn_thomas. (2023, June). *What is going on with the sudden increase of bot accounts on Reddit*? Reddit. https://www.reddit.com/r/OutOfTheLoop/comments/144hool/what_is_going_on_with_the_sudden_increase_of_bot/

Reddit. (2020). *Transparency report 2020*. reddit. https://www.redditinc.com/policies/transparency-report-2020-1

Roth, Y., & Pickles, N. (2020, May 18). *Bot or not? the facts about platform manipulation on Twitter*. Twitter. https://blog.twitter.com/en_us/topics/company/2020/bot-or-not

Tait, Amelia. (2019, March 26). *The Secret Trail of Money Behind Those Instagram Porn Bots*. Vice News. https://www.vice.com/en/article/mbzd84/the-secret-trail-of-money-behind-those-instagram-porn-bots

Uhlemann, Thomas. (2023, August 21). *A Bard's Tale – how fake AI bots try to install malware*. https://www.welivesecurity.com/en/scams/a-bards-tale-how-fake-ai-bots-try-to-install-malware/

Walsh, Dylan. (2023, June 12). *Study finds bot detection software isn't as accurate as it seems*. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/study-finds-bot-detection-software-isnt-accurate-it-seems