Artificial Humanity: Counteracting the Threat of Bot Networks on Social Media

POLICY ISSUE

Text-based social networks (such as, but not limited to Twitter) serve as powerful forums for public political discourse, increasingly becoming a scalable, digital agora - a central public space as understood by ancient Greek city-states, where both merchant activity, civic discussion and political debate take place. This space, however, is now under threat from social bots (Ferrara et al., 2016) that can masquerade as humans, seize control of this public space and create the illusion of overwhelming social proof behind any cause the operators chose to champion. Operators of these social networks have been shown to be often unaware, incapable or unwilling to deal with social bots. Given the threat this poses to democratic discourse, we recommend that governments be proactive in detecting and neutralizing such operations on their home turf.

TARGET POLICYMAKER/S

ICT bodies maintained by or affiliated with national governments.

RECOMMENDATIONS

Consult with public and civic bodies to extend existing policies and mechanisms to discourage impersonation of digital identities.

Create legal mechanisms for punitive action against operators of social bots found to be intentionally disseminating hate speech, misinformation, and political propaganda founded on such.

Create mechanisms to alert the public to the existence of social bot operations and invest in public education of a human-usable framework for the verification of digital identities online.

Create mechanisms amongst law enforcement bodies to actively engage with platforms and remove such operations at the source.

Invest in monitoring operations, ideally as a collaboration between government ICT bodies and journalists, that can utilize computer-aided analysis and human interpretation to regularly to detect social bot networks, especially with regards to the addition of artificial social weight to information and disruption of public discourse.

THE RESEARCH

THE ORIGIN OF THE SPECIES

The advent of bots stretch back to 1966 (Weizenbaum, 1966), where a simple program called ELIZA was successfully able to fool humans into believing that they were having a conversation with a real human being.

Technology has, of course, progressed. Today we see more capable software bots built into smartphones, tasked to perform many functions that secretaries once did. However useful, these technologies pose a threat when deployed on social media. As Ferrera et al. (2016) notes, "*the novel challenge brought by bots is the fact they can give the false impression that some piece of information, regardless of its accuracy, is highly popular and endorsed by many, exerting an influence against which we haven't yet developed antibodies. Our vulnerability makes it possible for a bot to acquire significant influence, even unintentionally. Sophisticated bots can generate personas that appear as credible followers, and thus are more difficult for both people and filtering algorithms to detect*".

These bots, portraying the impression of actual humans interacting, can collaborate as networks to subvert or shut down political discussion, disseminate misinformation and hate speech and amplify viral political propaganda in local niches. Bots have been documented threatening activists, attempting to swing elections and even engaging in debate with Donald Trump online . Some estimates place as many as 50 million Twitter accounts to be bots (Diresta et al., 2017) - an army of virtual humans more than twice the population of Sri Lanka.

While such activities have hitherto primarily been the worry of the West, we present evidence that social bots are invading the social media landscape of the Global South. Following the March 2018 Sinhala-Muslim mob violence in the Sri Lankan city of Kandy, Sri Lankan Twitter users began tweeting about large numbers of suspicious accounts that had suddenly started to follow them. A technical investigation, involving the user block lists of over 40,000 users, revealed several thousand such bots in the Sri Lankan twitter space (Hattotuwa, Wijeratne, Serrato, 2018).

An analysis of a small sample revealed that they all followed prominent personalities of some political authority on social media. Further analysis of some 200,000 tweets of those personalities revealed the logic behind the bot network: they tended to follow prominent social media users who tended to share articles online, thus potentially acting as sources of news and opinion for those following them. The social bots had Sinhala, Muslim and Tamil sounding names, representing Sri Lanka's three largest demographics. Many of the profiles were the default Twitter profile image, but many of the female accounts had photos lifted from public profiles of other individuals

A phenomenon similar to what was documented in Sri Lanka was reported from Thailand, Malaysia, China, Hong Kong, and other nations. Notably, Twitter did not appear to act on this knowledge (Russell, 2018). This is pattern is analogous to a spate of such accounts that appeared in the Myanmar Twittersphere after the 2017 Rakhine attacks. Given that the Myanmar bots began to engage in political conversation, we see anecdotal evidence (as presented in Hattotuwa, Wijeratne, Serrato, 2018) that there could be one or more actors in the South Asian space building a network of bots which may eventually be used for such purposes.

Some social bots are in principle innocent - news aggregators and autoresponders adopted by companies are two examples. However, analysis around the Boston Marathon Bombing has revealed such innocuous bots retweeting unverified content until it led to a spate of fake news. The potential damage to public perception is often difficult to quantify, given a lack of research establishing universal correlations between social media and offline political and economic behaviour. However, notable examples should provide some understanding of the degree of linkage between social media and the offline world.

In 2013, the Syrian Electronic Army allegedly hacked the Twitter account of the Associated Press, tweeting that two explosions had taken place in the White House and that President Obama had been injured. The market reacted immediately, with the Dow Jones industrial average plunging more than 100 points.

As the importance of social networks to political discourse in the Global South rises, propagandists and other actors may seek to control these discussions by unleashing armies of bots, with sufficiently human behaviour to fool the casual user - and many others may unwittingly have similar effects without any such intention to.

As posited before, platforms like Twitter do have rules against this type of activity. However, from the evidence given above, their enforcement is clearly lacking. It could be that the detection of these networks is difficult: global platforms may simply not be aware of botting due to lack of local context, language capacity and other factors. As a case in point, much of the research cited above is for the detection of social bots generating English context: many other languages may not have been dealt with at all.

Alternatively, platforms may simply not care to invest in the detection and eradication of the networks. This is a precarious situation for the Global South.

THE CHALLENGE OF SURVIVAL

Two problems present themselves in the context of the Global South: the first is that the majority popular social media are primarily US-based platforms, as they are set up and controlled by American citizens. Thus the Global South appears to have little of control required to enforce more stringent detection and elimination of bot threats from the side of the platforms.

The second problem is the investigation of bot networks. As cited previously, complicated mimicries have been observed, and are difficult to detect until after the damage is done. Documented methods of detecting and uncovering such social bots (Ratkiewicz et al., 2011; Costa, Yamaguchi, 2015; Davis et al., 2017) generally require some degree of technical expertise, investments in computing, and a keen awareness of the local context in which a social bot network may operate.

Thus, should a platform be willing to investigate such issues, they may face limitations in doing so due to lack of awareness of local languages and issues.

A key juncture will arrive when these such bot networks begin to more widely use modern technologies from the field of artificial intelligence. We have observed such bots in action, such as Microsoft Tay, which was programmed to learn from the people it interacted with, and interacted with a remarkable degree of realism, including internet slang (Vincent, 2018). Tay began to 'learn' patterns of racism, Holocaust denial and leanings towards authoritarianism. Its Japanese counterpart, Microsoft Rinna, charmed users with its schoolgirl personality (McKirdy,

2015). In both cases, it is worth noting how quickly these bots learned, and how realistic their recorded interactions are. Unless otherwise labelled, these would be almost impossible to tell apart from an actual human using Twitter. Similar bot operations, moreover, can easily be created with commercially available software.

Thus, we posit that the Global South is thus in need of a tripartite framework, constituting of a) institutions that can identify and flag these threats and produce tailored analyses b) a legal system that allows these analyses to be used to prevent malevolent actors from operating malicious bot networks within areas of their jurisdiction

c) mechanisms of alerting the public that use these spaces to identified bot networks and the biases that such networks might bring into relevant topics on social media channels.

These efforts should, where possible, facilitate public-private collaborations to overcome limitations of technical sophistication in government. A national body may be required for this task. We recommend that policymakers also periodically examine these measures to create and constantly update localized, enforceable rulesets to maintain a safer space for public discourse.

REFERENCES

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016, July 01). The Rise of Social Bots. Retrieved from https://cacm.acm.org/magazines/2016/7/204021-the-rise-of-social-bots/fulltext

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.

Diresta, R., Little, J., Morgan, J., Neudert, L. M., & Nimmo, B. (2017, November 02). The Bots That Are Changing Politics. Retrieved from https://motherboard.vice.com/en_us/article/mb37k4/twitter-facebook-google-bots-misinformation-changing-politics

Hattotuwa, S., Wijeratne, Y., & Serrato, R. (2018, June 09). Weaponising 280 characters: What 200,000 tweets and 4,000 bots tell us about state of Twitter in Sri Lanka. Retrieved from http://www.cpalanka.org/weaponising-280-characters-what-200000-tweets-and-4000-bots-tell-us-about-state-of-twitter-in-sri-lanka/

Russell, J. (2018, April 23). Twitter doesn't care that someone is building a bot army in Southeast Asia. Retrieved from https://techcrunch.com/2018/04/20/twitter-doesnt-care-that-someone-is-building-a-bot-army-in-southeast-asia/

Hattotuwa, S., & Wijeratne, Y. (2018, February 23). Namal Rajapaksa, bots and trolls: New contours of digital propaganda and online discourse in Sri Lanka. Retrieved from https://groundviews.org/2018/01/24/namal-rajapaksa-bots-and-trolls-new-contours-of-digital-propaganda-and-online-discourse-in-sri-lanka/

Ratkiewicz, J., Conover, M., Meiss, M. R., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Detecting and tracking political abuse in social media. *ICWSM*, *11*, 297-304.

Ferraz Costa, A., Yamaguchi, Y., Juci Machado Traina, A., Traina Jr, C., & Faloutsos, C. (2015, August). Rsc: Mining and modeling temporal activity in social media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 269-278). ACM.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*.

Vincent, J. (2016, March 24). Twitter taught Microsoft's friendly AI chatbot to be a racist asshole in less than a day. Retrieved from https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

McKirdy, A. (2015). Line's AI program captures hearts with lifelike personality. Retrieved from https://www.japantimes.co.jp/news/2015/08/06/business/tech/lines-ai-program-captures-hearts-lifelike-personality/#.WyDm7UiFOFE

Yudhanjaya Wijeratne | LIRNE*asia* | 12, Balcombe Place, Colombo 08, Sri Lanka | yudhanjaya@lirneasia.net.

Sanjana Hattotuwa | Center for Policy Alternatives | 6/5, Layards Road, Colombo 05, Sri Lanka

Raymond Serrato | Democracy Reporting International | Prinzessinnenstraße 30, 10969 Berlin, Germany

This work was carried out with financial support from the International Development Research Centre, Canada. The views expressed in this work are those of the creators and do not necessarily represent those of the International Development Research Centre, Canada or its Board of Governors.