Fighting online trolls with bots

January 11 2017, by Saiph Savage

Credit: AI-generated image (disclaimer)

The wonder of internet connectivity can turn into a horror show if the people who use online platforms decide that instead of connecting and communicating, they want to mock, insult, abuse, harass and even threaten each other. In online communities since at least the early 1990s, this has been called "trolling." More recently it has been called cyberbullying. It happens on many different websites and social media systems. Users have been fighting back for a while, and now the owners and managers of those online services are joining in.

The most recent addition to this effort comes from Twitch, one of a few increasingly popular platforms that allow gamers to play video games, stream their gameplay live online and type back and forth with people who want to watch them play. Players do this to show off their prowess (and in some cases make money). Game fans do this for entertainment or to learn new tips and tricks that can improve their own play.

Large, diverse groups of people engaging with each other online can yield interesting cooperation. For example, in one video game I helped build, people watching a stream could make comments that would actually give the player help, like slowing down or attacking enemies. But of the thousands of people tuning in daily to watch gamer Sebastian "Forsen" Fors play, for instance, at least some try to overwhelm or hijack the chat away from the subject of the game itself. This can be a mere nuisance, but can also become a serious problem, with racism, sexism and other prejudices coming to the fore in toxic and abusive comment threads.

In an effort to help its users fight trolling, Twitch has developed bots – software programs that can run automatically on its platform – to monitor discussions in its chats. At present, Twitch's bots alert the game's host, called the streamer, that someone has posted an offensive word. The streamer can then decide what action to take, such as blocking the user from the channel.

Beyond just helping individual streamers manage their audiences' behavior, this approach may be able to capitalize on the fact that online bots can help change people's behavior, as my own research has documented. For instance, a bot could approach people using racist language, question them about being racist and suggest other forms of interaction to change how people interact with others.

Using bots to affect humans

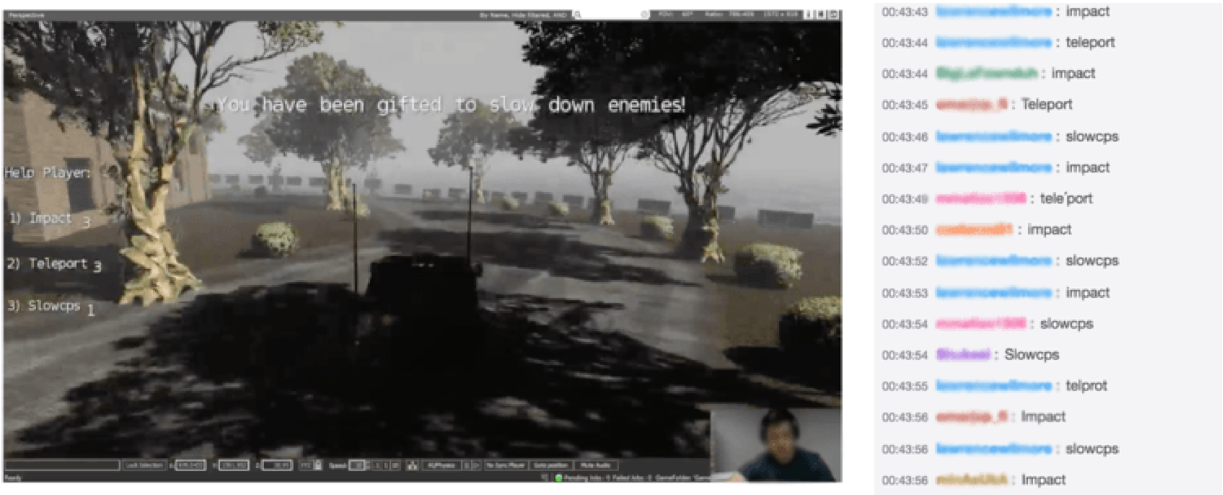
In 2015 I was part of a team that created a system that uses Twitter bots to do the activist work of recruiting humans to do social good for their community. We called it Botivist.

We used Botivist in an experiment to find out whether bots could recruit and make people contribute ideas about tackling corruption instead of just complaining about corruption. We set up the system to watch Twitter for people complaining about corruption in Latin America, identifying the keywords "corrupcion" and "impunidad," the Spanish words for "corruption" and "impunity."

When it noticed relevant tweets, Botivist would tweet in reply, asking questions like "How do we fight corruption in our cities?" and "What should we change personally to fight corruption?" Then it waited to see if the people replied, and what they said. Of those who engaged, Botivist asked follow-up questions and asked them to volunteer to help fight the problem they were complaining about.

We found that Botivist was able to encourage people to go beyond simply complaining about corruption, pushing them to offer ideas and engage with others sharing their concerns. Bots could change people's behavior! However, we also found that some individuals began debating whether – and how – bots should be involved in activism. But it nevertheless suggests that people who were comfortable engaging with bots online could be mobilized to work toward a solution, rather than just complaining about it.

Humans' reactions to bots' interventions matter, and inform how we design bots and what we tell them to do. In research at New York University in 2016, doctoral student Kevin Munger used Twitter bots to engage with people expressing racist views online. Calling out Twitter users for racist behavior ended up reducing those users' racist communications over time – if the bot doing the chastising appeared to be a white man with a large number of followers, two factors that conferred social status and power. If the bot had relatively few followers or was a black man, its interventions were not measurably successful.



You have been gifted to slow down enemies!

Help Player:
1) Impact 3
2) Teleport 3
3) Slowcps 1

```
00:43:43 [redacted] : impact
00:43:44 [redacted] : teleport
00:43:44 [redacted] : impact
00:43:45 [redacted] : Teleport
00:43:46 [redacted] : slowcps
00:43:47 [redacted] : impact
00:43:49 [redacted] : tele´port
00:43:50 [redacted] : impact
00:43:52 [redacted] : slowcps
00:43:53 [redacted] : impact
00:43:54 [redacted] : slowcps
00:43:54 [redacted] : Slowcps
00:43:55 [redacted] : telprot
00:43:56 [redacted] : Impact
00:43:56 [redacted] : slowcps
00:43:56 [redacted] : Impact
```

When spectators get involved, they can help a player out. Credit: Saiph Savage, CC BY-ND

Raising additional questions

Bots' abilities to affect how people act toward each other online brings up important issues our society needs to address. A key question is: What types of behaviors should bots encourage or discourage?

It's relatively benign for bots to notify humans about specifically hateful or dangerous words – and let the humans decide what to do about it. Twitch lets streamers decide for themselves whether they want to use the bots, as well as what (if anything) to do if the bot alerts them to a problem. Users' decisions not to use the bots include both technological factors and concerns about comments. In conversations I have seen among Twitch streamers, some have described disabling them for causing interference with browser add-ons they already use to manage their audience chat space. Other streamers have disabled the bots because they feel bots hinder audience participation.

But it could be alarming if we ask bots to influence people's free expression of genuine feelings or thoughts. Should bots monitor language use on all online platforms? What should these "bot police" look out for? How should the bots – which is to say, how should the people who design the bots – handle those Twitch streamers who appear to enjoy engaging with trolls?

One Twitch streamer posted a positive view of trolls on Reddit:

"…lmfao! Trolls make it interesting […] I sometimes troll back if I'm in a really good mood […] I get similar comments all of the time…sometimes I laugh hysterically and lose focus because I'm tickled…"

Other streamers even enjoy sharing their witty replies to trolls:

"…My favorite was someone telling me in Rocket League "I hope every one of your followers unfollows you after that match." My response was "My mom would never do that!" Lol…"

What about streamers who actually want to make racist or sexist comments to their audiences? What if their audiences respond positively to those remarks? Should a bot monitor a player's behavior on his own channel against standards set by someone else, such as the platform's administrators? And what language should the bots watch for – racism, perhaps, but what about ideas that are merely unpopular, rather than socially damaging?

At present, we don't have ways of thinking about, talking about or deciding on these balancing acts of freedom of expression and association online. In the offline world, people are free to say racist things to willing audiences, but suffer social consequences if they do so around people who object. As bots become more able to participate in, and exert influence on, our human interactions, we'll need to decide who sets the standards and how, as well as who enforces them, in online communities.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

Citation: Fighting online trolls with bots (2017, January 11) retrieved 3 May 2024 from https://phys.org/news/2017-01-online-trolls-bots.html