Ray of light

Actor, activist, funny woman: Tracee Ellis Ross looks on the bright side of life
The internet opened the floodgates for extremism and a rising tide of hate speech.

Can tech help close them again?

**Trigger warning**

Simon Parkin meets the online ‘hate detectives’

Illustration by The Heads of State
Last month, HatleId identified three forms of Cove’s hate speech - anti-Chinese, antisemitic, Islamophobic: A health threat is being weaponised

When Coven realised the children in the Afghan training camp were being used for suicide attacks, his first instinct was to exact revenge on the people who had exploited him. “But I only had a knife, a gnog,” he says.

“And I knew that I couldn’t tell anyone I wanted out.”

He stepped outside the gate and confronted some of his options.

He pulled out his box of pick-up trucks with whom he had talked a few times. “The men did not share his first language, but Coven gambled. He pulled out $50 and moved at the request, asking if he could fit a ride in the Shades on the border between Pakistan and Afghanistan. The driver agreed. Coven climbed into the passenger seat. "I didn’t even look back," he recalls.

"It’s a serious health threat," the program reflects a true indication of the prevalence of online hatred.

It offers a groundbreaking tool. Likewise, Williams speculates a date range, then plots out potential target groups: jews, homosexuals, women, and so on. His work is by far the most prevalent form of hate speech on the internet today and it is ever growing.
the police. “People are fearful of secondary victimisation,” Williams says.

As domestic internet use became more commonplace, Williams noticed
the hate speech he encountered on the streets reflected online. The
difference was that it was there for everyone to witness. Fellow academics
were initially sceptical of his preoccupation with online behaviour, but by
2011 “everyone knew hate speech was the key problem of the internet”. That
year, Williams received a lottery grant of more than half a million pounds to
accelerate his research.

Every social media platform represents a torrent of information too
deep and wide to sift by hand. Williams and his team began by taking a
random sample of 4,000 tweets from a dataset of 200,000. The trove was
then handed to four police officers, trained to recognise racial tensions,
who each evaluated whether every tweet was discriminatory. If three of
the four officers concurred, the tweet was classified as hate speech. Over a
four-week period, the officers identified around 600 tweets they deemed
discriminatory, data that formed the gold standard by which the AI would
test if a message was “malignant” or “benign”.

On the afternoon of 22 May 2013, when fusilier Lee Rigby was killed by two
Islamist converts in Woolwich, London, the software had its first live test.
Within 60 minutes of the attack, Williams and his team began harvesting
tweets that used the keyword “Woolwich”. As the software sifted the data,
the team was able to examine the drivers and inhibitors of hate speech, and
identify accounts spreading anti-Muslim rhetoric. The team found that hate
speech peaked for 24-48 hours, and then rapidly fell, while the baseline of
online hate remained elevated for several months. Astonishingly, this was
one of the first times a link between terror attacks and online hate speech
had been demonstrated. And importantly, an increase in localised hate speech
both anticipated the attack and, in the aftermath, shadowed it, showing that
it might be possible to predict real world attacks.

The data fascinated social scientists, but Williams believed it was more
than interesting: it could have a practical application in helping counter
these narratives. In 2017, he began a pilot scheme with the national online
hate crime hub, which was set up to coordinate reporting into this area. It
now uses the HateLab dashboard to gauge ebbs and flows in the targeting of
particular groups, as well as nuances in local tensions. This information can
then inform operational decisions, helping direct frontline police work.
There are obvious privacy concerns, and HateLab must comply with
data protection regulations. The platform depends on the willingness of
Twitter to make its data available to third-party applications. (Facebook
closed down open access in 2018, so independent organisations cannot
screen its posts.) Twitter shares data on the proviso that HateLab does not
identify individual accounts via its dashboard. “In that sense, we can only
provide the 10,000ft view,” Williams says. The dashboard can highlight
patterns, target groups and geographical hotspots - but connecting with
individuals is outside its remit.

**The AI showed a rise in local hate speech before the murder of Lee Rigby. Was this a key to predicting future attacks?**

Meanwhile, Qadir and the other former extremists working alongside
Moonshot recognise the power that hate speech can have, and know
firsthand that a conversation can steer someone down a more positive path.
“You can only change people if you can reach them via conversation,” he
tells me. “Violent extremists do this very cleverly, and evidence shows that
it works for them, so I based all my programmes on this concept. You have
to engage and create conversations, but direct them positively - allow for
grievances to be heard and discussed.”

Since Moonshot was founded, there has been a radical shift in the
perception of technology’s role when it comes to extremist terrorism. “Five
years ago, there were still people inside the government who thought tech
was for the kids,” Frennet says. “There was a sense that it was almost amusing
that terrorists were on the Internet. You don’t get that any more. Likewise, five
years ago there were some great organisations doing great work on the violent
far-right, but again it was almost seen as niche. That’s no longer the case.”

---

The Guardian Weekend | 02 May 2020 19