Ray of light

Actor, activist, funny woman: Tracee Ellis Ross looks on the bright side of life

Time out: adventures of a sports fan in lockdown
Speak no evil: the online war against hate speech
Romesh Ranganathan: burpees in my kitchen

The Guardian
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
When Qadir realised the children in the Afghan training camp were being mistreated, he first instinct was to exact revenge. "But I knew that I couldn't kill anyone I wanted."

"I stepped outside the gate," Qadir continued, "and took a pick-up truck with whom he talked a few times. The men did not share a first language, but Qadir gambled. He pulled out a gun and waved it at the men, asking if he could take a ride. With his long braided hair and tall stature, Qadir climbed into the back seat. "I didn't even collect my bag," he recalls.

"Qadir said he tried to talk to the Afghan government and the United Nations. He told them about the situation and the children who were being mistreated. But he knew that it was a long shot. Qadir had no way to know if anyone would listen or if they even cared."

The last month, HateLabel identified three forms of  

"Hate speech - anti-Chinese, antisemitic, Islamophobic: A health threat is being weaponised"

When Qadir realised the children in the Afghan training camp were being mistreated, he first instinct was to exact revenge. "But I knew that I couldn't kill anyone I wanted."

"I stepped outside the gate," Qadir continued, "and took a pick-up truck with whom he talked a few times. The men did not share a first language, but Qadir gambled. He pulled out a gun and waved it at the men, asking if he could take a ride. With his long braided hair and tall stature, Qadir climbed into the back seat. "I didn't even collect my bag," he recalls.

"Qadir said he tried to talk to the Afghan government and the United Nations. He told them about the situation and the children who were being mistreated. But he knew that it was a long shot. Qadir had no way to know if anyone would listen or if they even cared."

The last month, HateLabel identified three forms of  

"Hate speech - anti-Chinese, antisemitic, Islamophobic: A health threat is being weaponised"

When Qadir realised the children in the Afghan training camp were being mistreated, he first instinct was to exact revenge. "But I knew that I couldn't kill anyone I wanted."

"I stepped outside the gate," Qadir continued, "and took a pick-up truck with whom he talked a few times. The men did not share a first language, but Qadir gambled. He pulled out a gun and waved it at the men, asking if he could take a ride. With his long braided hair and tall stature, Qadir climbed into the back seat. "I didn't even collect my bag," he recalls.

"Qadir said he tried to talk to the Afghan government and the United Nations. He told them about the situation and the children who were being mistreated. But he knew that it was a long shot. Qadir had no way to know if anyone would listen or if they even cared."

The last month, HateLabel identified three forms of  

"Hate speech - anti-Chinese, antisemitic, Islamophobic: A health threat is being weaponised"

When Qadir realised the children in the Afghan training camp were being mistreated, he first instinct was to exact revenge. "But I knew that I couldn't kill anyone I wanted."

"I stepped outside the gate," Qadir continued, "and took a pick-up truck with whom he talked a few times. The men did not share a first language, but Qadir gambled. He pulled out a gun and waved it at the men, asking if he could take a ride. With his long braided hair and tall stature, Qadir climbed into the back seat. "I didn't even collect my bag," he recalls.

"Qadir said he tried to talk to the Afghan government and the United Nations. He told them about the situation and the children who were being mistreated. But he knew that it was a long shot. Qadir had no way to know if anyone would listen or if they even cared."

The last month, HateLabel identified three forms of  

"Hate speech - anti-Chinese, antisemitic, Islamophobic: A health threat is being weaponised"
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 tweets were 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 ebb 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.”

Vidhya Ramalingam and Ross Frennet set up Moonshot, a research initiative, after the murder of 77 people by Anders Breivik in Norway in 2011.