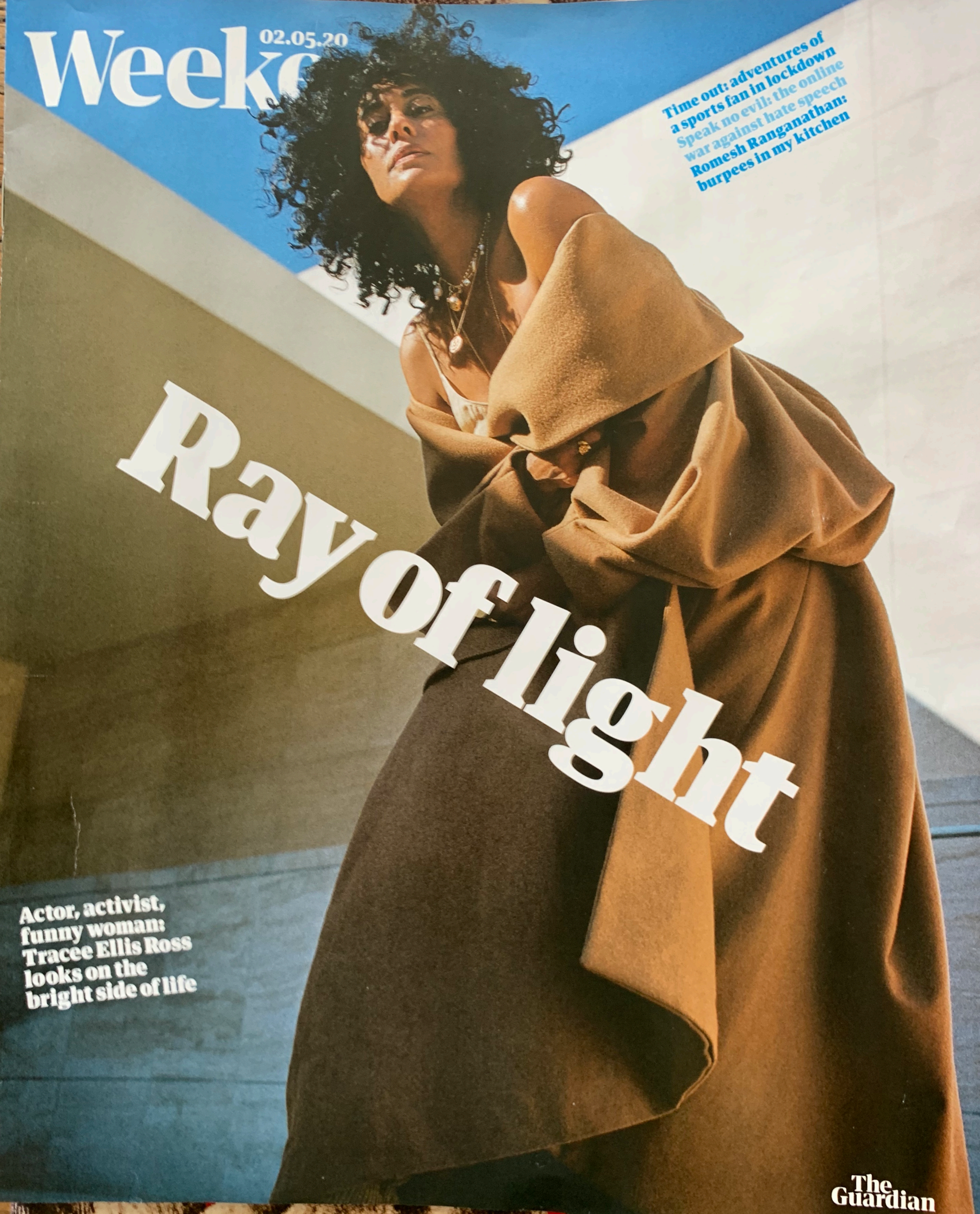
Weekend 02.05.20



Time out: adventures of a sports fan in lockdown

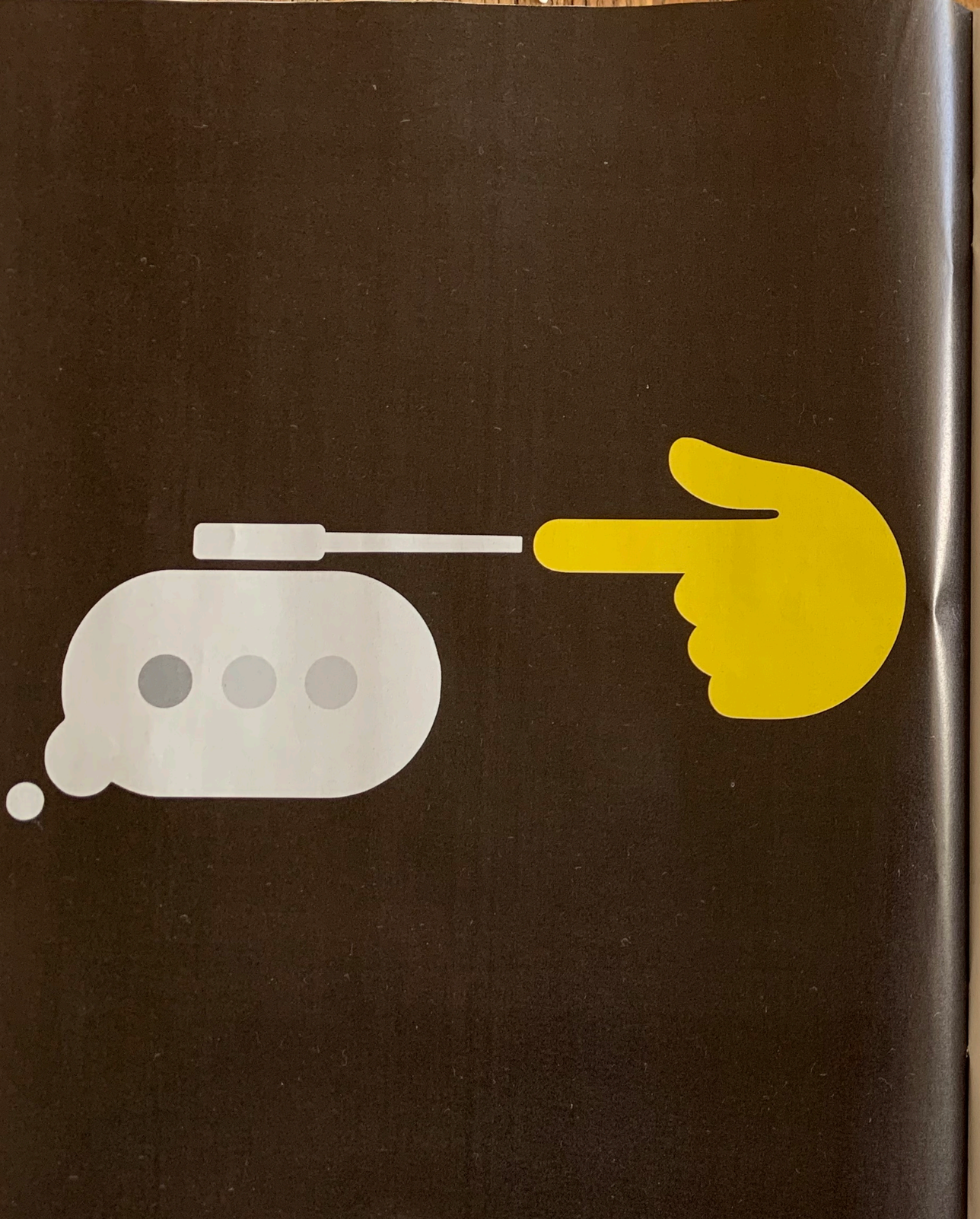
Speak no evil: the online war against hate speech

Romesh Ranganathan: burpees in my kitchen

Ray of light

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

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

In the winter of 2002, nine months before Hanif Qadir unpacked his bag at a terrorist training camp in Afghanistan, a group of men walked into the London MOT testing centre he owned with his brothers. They were collecting money for civilians caught up in the US invasion of Afghanistan; hundreds of children had been orphaned by indiscriminate bombing, the men claimed. Could he help? The appeal resonated with Qadir, who had lost his father when he was seven. He made a donation.

The men returned regularly. Each time, they asked for more money, before gradually changing the subject to Qadir's faith. Eventually they invited him to a meeting at a local house to discuss the war in Afghanistan more freely. "I felt they were sincere and genuine," Qadir recalls. At the meeting, the men encouraged Qadir to visit websites that claimed to show photographic evidence of violence against Afghan civilians by western troops.

Qadir browsed hundreds of distressing images, among them scores of orphans, each accompanied by extended captions that described the way in which the child's family had been killed. One girl's story has remained with him. The website claimed she had lost 21 members of her family to a "stray" US missile. The caption explained it had taken locals three days to scrape their remains from the walls of the girl's home. The more he saw, the closer Qadir became to the men who were, unbeknown to him, recruiters for al-Qaida.

Qadir grew up in Thornaby-on-Tees, a small town in North Yorkshire. After his father died, he had disengaged from school, leaving at 14 and moving to London. After a few odd jobs, he founded a business with his brothers, buying, repairing and selling cars. By the early 2000s, the business was profitable enough that he was able to donate generously to charitable community causes, a reputation that, he believes, led the recruiters to his door.

The suggestion that Qadir travel to Afghanistan was seeded gently. "When a person is radicalised they become suggestible," he tells me. "We discussed that, in order to prevent more loss of life, we needed to be prepared to fight." On 2 December 2002, he flew to Islamabad in Pakistan. A few days later, he crossed the border into Afghanistan.

Soon after he arrived at a training camp, Qadir saw a man measuring up children who lived there. "I thought they were being tailored for new clothes," he recalls. Then he heard one of the leaders telling the children they would soon be reunited with their dead parents. They were being fitted for suicide vests. "I felt sick and angry," he says. "I wanted to walk away."

But in the middle of a desert compound patrolled by armed guards, any attempt to defect could be fatal. Qadir was trapped. "I knew that if I asked to leave things would end badly." He had to think carefully.

In 2002, when Qadir was being radicalised, the internet was not yet ubiquitous. There was no Twitter, no Facebook; websites looking to groom people into supporting extremist causes were obscure. Two decades later, the digital landscape has been transformed. As the All-Party Parliamentary Group on Hate Crime wrote last year, the internet has become a "key breeding ground" for extremism and hate speech - emboldened by the increasing ease of dissemination, anonymity and, thanks to outdated legislation, a lack of meaningful consequences.

Perpetrators of terrorist attacks now routinely leave online statements or manifestos to justify their actions, hoping their words might encourage others. The 28-year-old gunman who killed 51 mosque-goers in Christchurch, New Zealand, last year posted a 73-page white nationalist rant to the fringe web forum 8chan and livestreamed the attack on Facebook.

But now, just as Facebook and Twitter have become the prodigious muck-spreaders of our age, a handful of clandestine startups are using technology to stem the flow. Moonshot, whose office is at a secret location in London, is, at five years old, a veteran in this emerging industry. Its premises have the feel of a typical Silicon Valley operation: distressed floorboards, glass-fronted offices, beanbags by an open fireplace, exposed brickwork, a snug for breathers. There are a few clues that the company's business - using technology to disrupt violent extremism - is different from that of the fitness app developers, social media influencers and virtual reality speculators with whom it shares an aesthetic. The posters are not vintage prints but disquieting infographics revealing, for example, that after 22 people were shot dead in an El Paso Walmart last August, there was an 82% rise in the Google search term "how to murder Mexicans". There is also a bomb-proof door.

Cofounder Vidhya Ramalingam set up the EU's first intergovernmental research initiative to investigate far-right terrorism in the aftermath of the 2011 murder of 77 people by Anders Breivik in Norway. She describes Moonshot's work as "experimental programming". The company employs 50 people, and uses a mixture of software and human judgment to identify individuals on the internet who, like Qadir, appear interested in extremist propaganda. They then attempt to serve them counter-messaging.

The technology uses a "database of indicators of risk". An individual is awarded "risk points" according to their online behaviour. You score one point for showing curiosity about the Ku Klux Klan or National Socialist Movement. Activity that indicates sympathy with a violent movement or ideology (eg Googling "white pride worldwide") earns three points, while showing a desire to join, send money to, or commit acts on behalf of a violent extremist group or individual earns six.

Home Office initiatives such as Prevent have traditionally focused on training teachers and other leaders to identify people likely to be drawn to violent extremism within their communities - but these methods risk introducing discriminatory practices. "In France, for example, there were posters telling people their sons might be at risk of violent extremism if they grow a beard, start speaking Arabic or stop eating baguettes," explains Ross Frenett, Moonshot's cofounder. "That is obvious bullshit."

By contrast, Frenett says, if someone makes a post glorifying Hitler, or calls for genocide against Muslims, there is a high degree of certainty that they fall into a high-risk category. "They've essentially tattooed a swastika to their forehead in the online space," he says. "So our level of confidence when identifying individuals who are vulnerable to radicalisation is way higher online than it could ever be offline. And it sidesteps some of the discriminatory, stigmatising practices we've seen in an offline setting."

Moonshot, founded in September 2015, is a for-profit company that earns its income from government contracts in the UK, US, Canada, Australia and across western Europe. It does not limit its work to any particular strain of radicalism; in addition to the far-right and jihadism, Moonshot's work covers everything from Buddhist extremism in south Asia, to Hindu nationalism and "incel" terrorism in Canada.

At the broadest level, Moonshot runs what it refers to as "redirection material" - advertising that is designed to "get in front of" extremist material

Last month, HateLab identified three forms of Covi

in Google's search results. Google has granted Moonshot dispensation to advertise against banned search terms such as "join Isis". If a user clicks on one of Moonshot's camouflaged results, they are taken to, for example, a mental health website with relevant downloadable guides and a chat option. (These sites are run by partnered mental health organisations and groups that have experience dealing with gang violence. As Frenett puts it, they have "appropriate risk protocols, and connections with law enforcement, should they be required".) So long as the search terms are carefully calibrated (advertising against "white power" is useless, Frenett explains, as you end up competing with power-tool companies) this can be an effective first contact.

Success is measured in much the same way as any company seeking to advertise on Google, via click conversions. ("We pay for advertising just like any commercial advertiser does," Ramalingam says. "We don't get special rates. I wish we had a better story on that front.") A key metric is "search impression share", which records the amount of time your at-risk audience

Hanif Qadir has used his experience of being radicalised to help others

AP PHOTO

saw the ad. "We've had campaigns that have run with only 50%, and that's not good enough," Ramalingam says. "So we work hard to get that up to 98% where possible." For this reason, as well as mental health practitioners and ex-police officers, Moonshot also employs marketers. "Most of our work is analytics, marketing and social work," Frenett says. "It just happens to be marketing, analytics and social work related to terrorism."

Occasionally the company will identify an individual who is too high risk for their interventions. "That's where, depending on the country we're working in, we refer a user to the police," Frenett says. In Australia, for example, Moonshot identified someone at the top of a network of around 200 at-risk individuals considered "so risky we couldn't intervene". A few days later, the local police arrested the man, who was subsequently convicted on terror charges.

There are deeper kinds of intervention. One of Moonshot's advertisements for, say, "bomb manuals" will take the searcher to a WhatsApp chat manned by a specialist trained in deradicalisation techniques. The company may also identify someone on a particular social media platform openly espousing pro-extremist or pro-terrorist views. Then a trained social worker, typically from a charitable partner organisation, contacts that individual via Twitter direct message or Facebook Messenger. Choosing the right person to make this kind of contact, which may be perceived as invasive, is essential. In many cases, the right person is a former extremist - someone like Hanif Qadir.

car business and convinced his brothers to sell up. "I just wanted to stay at home with my children."

After a period of recuperation, he and his brothers opened a gym in a disused nightclub, which became a place where local youths, many of them young Muslims, would congregate. "We'd talk," he says. "I'd ask them questions about Afghanistan. I saw a lot of anger and questioning. It was clear to me that all it would take is for one person to manipulate them emotionally and they would get straight on a plane to fight. Or maybe they would do something here."

Eager to communicate this to someone in a position of power, Qadir started attending local council meetings. A police inspector, Ian Larnder, took him for a coffee, hoping to better understand why this former mechanic seemed so passionate about the subject. "Until then, I had told nobody about what had happened," Qadir recalls. "Ian was the first person I opened up to." A week later, Larnder was appointed to the police's national community tension team. He took Qadir with him to speak to forces around the country about his experiences.

Today, with a number of other former extremists, Qadir works with Moonshot, where he provides training for online interventions. "The skill is in finding out what has raised a person's interest in extremist ideology," he explains. "You can't redirect a person until you understand this. It's no good asking something so broad as: 'What do you think about what is happening in India?' It has to be specific and personable. So instead you might say: 'Is it permissible to seek revenge for the loss of a loved one?'"

This sort of broad line of questioning - and the fact that an anonymous dialogue might tail off, without scope for any follow-up - can seem frustratingly opaque for anyone trying to measure Moonshot's success. It's a criticism the company is used to fielding. "The struggle with preventive work is that, very often, it's unscientific and we have to ask people to take it on trust," Frenett says. "It's easy for a military contractor to come in and say, 'I installed a big, high fence and a man with a gun and that reduced terrorism.' Likewise, the army can come along and state: 'We killed 200 Taliban this week.'

"But it's much harder to say, 'OK. We invested $1m here and we prevented this much terrorism.' Our long-term aim is to start to change that calculation. Then we'll be able to say: 'If one dollar in every 100 spent on military hardware went towards targeted, community-focused preventive work it would be better value - and probably better for the world.'"

In the corner of a chilly room at the end of a corridor in Cardiff University's Glamorgan Building, a flood of racial slurs, misogyny, antisemitism and far-right slogans flows across a PC screen. "Imagine you had a crystal ball in which you could watch someone perpetrating every hate crime as it occurred somewhere out there, on the streets," explains Matthew Williams, director of HateLab. "That's what you're

19 hate speech - anti-Chinese, antisemitic, Islamophobic: 'A health threat is being weaponised'

When Qadir realised the children in the Afghan training camp were being measured for suicide vests, his first instinct was to exact revenge on the people who had manipulated him. "But I only had a knife, no gun," he says. "And I knew that I couldn't tell anyone I wanted out."

He stepped outside the gate of the camp to consider his options. There he spotted the driver of a pick-up truck with whom he had talked a few times. The men did not share a first language, but Qadir gambled. He pulled out £50 and waved it at the man, asking if he could hitch a ride to Turkham, on the border between Pakistan and Afghanistan. The driver nodded. Qadir climbed into the passenger seat. "I didn't even collect my bag," he recalls.

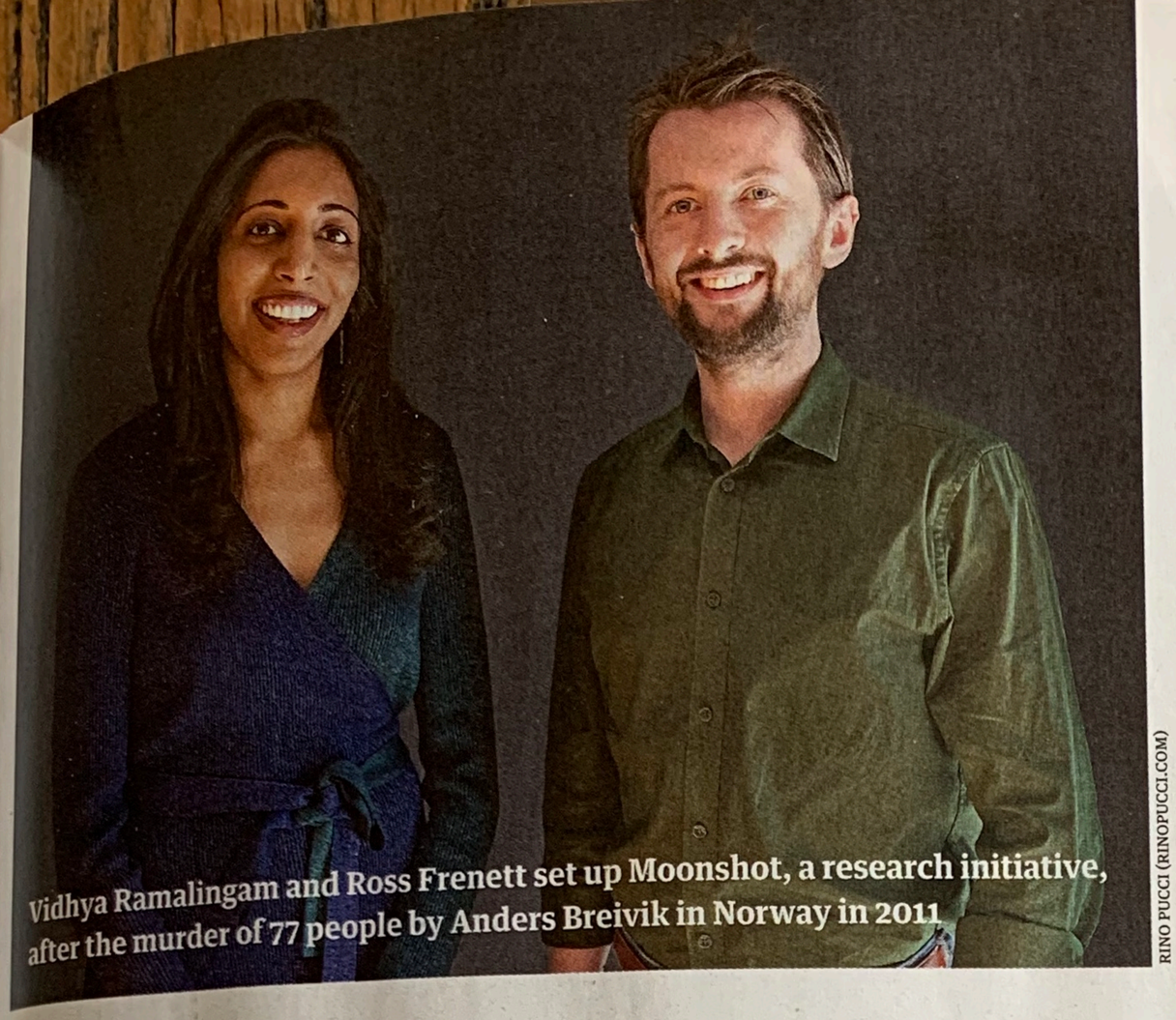
Qadir says he cried during the flight to London. "I kept asking myself: 'What the hell have I done?'" When he arrived home, he and his brothers attempted to find his recruiters, but they had disappeared; the word was that they had moved to Manchester. Qadir decided he no longer wanted to run the looking at here, except the hate is happening online."

While Moonshot and Qadir intervene with individuals who are vulnerable to extremism, HateLab's aim is to provide a more accurate picture of hate speech across the internet. It is, Williams says, the first platform to use AI to detect online hate speech in real time and at scale.

Online hatred is so commonplace that the majority of incidents go unreported. According to British government data, 1,605 hate crimes occurred online between 2017 and 2018, a 40% increase on the previous year. But the Home Office admits this figure is probably a gross underestimate.

"Unlike the police, we don't have to wait for a victim to file a report," Williams says. "The program reflects a true indication of the prevalence of online hatred."

It offers a granular indication, too. Williams specifies a date range, then picks from a filter of potential target groups: Jews, homosexuals, women, and so on (misogyny is by far the most prevalent form of hate speech on »



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

RINO PUCCI (RINOPUCCI.COM)

Twitter, he says). He selects "anti-Muslim" and a heat map of the UK lights up in red blotches showing geographical hotspots. Elsewhere, it reports the average number of hateful posts per minute and the peak times of day (hate speech, the group has found, is most prevalent during the daily commute, when people read and react to the day's news).

A word cloud indicates the most-used anti-Muslim slurs, while a spiderweb visualises a network of perpetrators, identifying the "thought leaders" who are generating the most retweets, and how they are linked, via online accounts. "HateLab gives situational awareness to hate speech on Twitter at any given time," Williams says.

Early last month, HateLab identified three forms of coronavirus-related hate speech: anti-Chinese or Asian; antisemitic, focused on conspiracy theories; and Islamophobic, focused on accusations of profiteering. "What we are seeing is a threat to health being weaponised to justify targeting minority groups, no matter how illogical the connections may seem," Williams explains.

(Moonshot has monitored similar rises in hate speech targeting Chinese nationals. The hashtag #ChinaLiedPeopleDied was tweeted 65,895 times in March, while #coronavirustruth, implying that the pandemic is a hoax, was used 77,548 times. The company also picked up tweets showing old videos of Muslim men leaving mosques accompanied by text claiming the footage was filmed during quarantine, a seemingly deliberate attempt to create anti-Muslim sentiment.)

Williams, author of a forthcoming book titled The Science Of Hate, is a professor of criminology at Cardiff, but his interest in the field is not purely academic. In 1998, he travelled to London with friends to celebrate a birthday. At some point during the evening, he stepped out of the gay bar in which the group was drinking. Three young men approached. One asked if Williams had a light. As he handed over his Zippo, the man punched him in the face. Williams returned to his friends but said nothing, fearing that they would want to retaliate. Eventually, one of them noticed blood on his teeth and urged him to report the attack. "I said no," Williams recalls. "At that time my parents didn't know I was gay. My siblings didn't know, and neither did most people from my town. I didn't want to come out to the police."

But Williams returned to Wales a changed person. "Any attack on your identity has a profoundly destabilising effect," he says. "I became angry and depressed. I modified my behaviour. I stopped holding my boyfriend's hand. I still won't show affection in public." He was not alone in failing to report his attackers; based on the combined 2015/16 to 2017/18 Crime Survey for England and Wales, only 53% of hate crime incidents came to the attention of 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," Frenett 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." ■