THE CHAOS MACHINE

THE INSIDE STORY OF HOW SOCIAL MEDIA REWIRED OUR MINDS AND OUR WORLD

MAX FISHER

Little, Brown and Company
New York • Boston • London

reject. The problem, in this experiment, wasn't ignorance or lack of news literacy. Social media, by bombarding users with fast-moving social stimuli, pushed them to rely on quick-twitch social intuition over deliberative reason. All people contain the capacity for both, as well as the potential for the former to overwhelm the latter, which is often how misinformation spreads. And platforms compound the effect by framing all news and information within high-stakes social contexts.

Politicians were adapting to this order. Matt Gaetz, a newly elected congressman from Florida, tweeted that shadowy powers were paying refugees to "storm" the border and disrupt the midterm elections and that Jewish philanthropist George Soros might be responsible. He was retweeted more than 30,000 times. The Russians weren't the problem anymore.

4. *A World Going Mad*

CHASLOT, STILL IN France, decided to repeat the tracking experiment he'd run on the American election, this time on the four-candidate presidential race at home. As before, YouTube's algorithm, he found, heavily favored the candidates at the extremes: the far-right Marine Le Pen and the far-left Jean-Luc Mélenchon. A new truism of politics was emerging: social media elevated anti-establishment politicians conversant with exaggerated moral-emotional language. Mélenchon, though unpopular with voters, won millions of views on YouTube, where his most dedicated fans seemed to congregate.

This had started as a positive: the internet offered political outsiders a way around the mainstream outlets that shunned them. As those candidates' grassroots supporters spent disproportionate time on YouTube, the system learned to push users to those videos, creating more fans, driving up watch time further. But thanks to the preferences of the algorithms for extreme and divisive content, it was mostly fringe radicals who benefited, and not candidates across the spectrum.

Backed by a handful of fellow researchers, Chaslot brought his findings on the American and French elections to *The Guardian*, resulting in an explosive report that offered seeming evidence of a long-suspected threat to global political stability. YouTube disputed "the methodology, data and, most importantly, the conclusions" of the research. Chaslot had made no secret that his conclusions were rough estimates, using thousands of datapoints to infer the algorithm's billions of daily decisions. But the findings were so consistent, he thought, and so consistently alarming, wouldn't the company want to look into it? Or share the internal data that could, in theory, clear all this up? And he was hardly alone. Over the following years, with the company stonewalling, an entire field of researchers published one set of findings after another, produced through ever more sophisticated methods, that not only supported Chaslot's results but suggested the reality was substantially worse than even he had feared.

Throughout, YouTube held to a consistent strategy, much like the one DiResta had described: deny, discredit, and antagonize. In response to the piece that ran in *The Guardian*, a spokesperson said, "Our only conclusion is that *The Guardian* is attempting to shoehorn research, data, and their incorrect conclusions into a common narrative about the role of technology in last year's election." This became a pattern. Time after time, the company's reps would respond to each new discovery by calling the evidence meaningless or wrong, digging in for long, often hostile exchanges. Then, once a major story ran, YouTube, in a paradoxical turnabout, would put out a statement insisting it had already fixed issues that, only weeks earlier, it had dismissed as nonexistent. In Chaslot's case, the company also sought to portray him as untrustworthy, motivated by a desire to embarrass the company in retaliation for having fired him for poor performance. But this could not explain why he had initially tried to put YouTube behind him, researching the platform only after seeing its harms firsthand years later, nor why he had initially taken his findings directly and privately to YouTube.

"That's the routine. I can laugh about it because, by changing things, they recognize that I was right," Chaslot said, though his voice was

suffused with a sadness over his former employer's public disavowals that, years later, still stung. "But when I was in the middle of it, they put such pressure on me. That was really frustrating."

It was a puzzling strategy, especially just as lawmakers had begun taking notice of social media's harms. Shortly after YouTube sent *The Guardian* its confrontational statement but before the newspaper went to press, the Senate Intelligence Committee sent Google a letter demanding that the company articulate its plan for preventing bad actors from manipulating YouTube's algorithm. YouTube asked to "update" its statement to *The Guardian*, replacing the vitriol with pledges to combat misinformation and praise for the newspaper's "work to shine a spotlight on this challenging issue."

Meanwhile, just as Chaslot joined DiResta and others in the public struggle to understand Silicon Valley's undue influence, William Brady and Molly Crockett, the psychologist and neuroscientist, achieved a momentous breakthrough in that effort. They had spent months synthesizing reams of newly available data, behavioral research, and their own investigations. It was like fitting together the pieces of a puzzle that, once assembled, revealed what may still be the most complete framework for understanding social media's effect on society.

The platforms, they concluded, were reshaping not just online behavior but underlying social impulses, and not just individually but collectively, potentially altering the nature of "civic engagement and activism, political polarization, propaganda and disinformation." They called it the MAD model, for the three forces rewiring people's minds. Motivation: the instincts and habits hijacked by the mechanics of social media platforms. Attention: users' focus manipulated to distort their perceptions of social cues and mores. Design: platforms that had been constructed in ways that train and incentivize certain behaviors.

The first stage of their findings had to do with how people perceive moral-emotional words. When Brady first found that such words travel further online, it had stood to reason that they draw attention because they

usually describe something dramatic. Brady decided to test this. He and two other scholars showed participants a fake social media stream, tracking what captured their attention as they scrolled. Moral-emotional words, they found, overrode people's attention almost regardless of context. If a boring statement with moral-emotional words and an exciting statement without them both appeared on screen, users were drawn to the former. Subjects actively focusing on something lost their concentration if a moral-emotional word so much as flashed elsewhere on the screen. Other sorts of flashing words did not produce the same effect.

When they reran the experiment with real tweets, they got the same results: the more moral-emotional words in a post, the more twitches of attention it won. Those posts also consistently had more shares. If you tweeted "The quick brown fox jumps over the lazy dog" and "The quick brown fox jumps over the liar dog," the latter would, from that one moral-emotional word, get more eyeballs and more shares. Tweet "The good hero fox slams the liar enemy dog," and you might be president by nightfall.

The digital-attention economy amplifies the social impact of this dynamic exponentially. Remember that the number of seconds in your day never changes. The amount of social media content competing for those seconds, however, doubles every year or so, depending on how you measure it. Imagine, for instance, that your network produces 200 posts per day, of which you have time to read 100. Because of the platforms' tilt, you will see the most moral-emotional half of your feed. Next year, when 200 doubles to 400, you see the most moral-emotional quarter. The year after that, the most moral-emotional eighth. Over time, your impression of your own community becomes radically more moralizing, aggrandizing, and outraged — and so do you. At the same time, less innately engaging forms of content — truth, appeals to the greater good, appeals to tolerance — become more and more outmatched. Like stars over Times Square.

Stage two in social media's distorting influence, according to the

MAD model, is something called internalization. Users who chased the platforms' incentives received immediate, high-volume social rewards: likes and shares. As psychologists have known since Pavlov, when you are repeatedly rewarded for a behavior, you learn a compulsion to repeat it. As you are trained to turn all discussions into matters of high outrage, to express disgust with out-groups, to assert the superiority of your in-group, you will eventually shift from doing it for external rewards to doing it simply because you want to do it. The drive comes from within. Your nature has been changed.

Brady and Crockett proved this in two experiments. In one, when users who expressed outrage were rewarded with likes and shares, they became likelier to express outrage in the future — and likelier to *feel* outraged. The effect held even for subjects who had earlier expressed an aversion to online anger. An otherwise sweet and tolerant person who, in a moment of weakness, sent a Democrat-bashing tweet that went viral would become instantly likelier to send more, first to chase the high, but soon because she had become, in her heart, more hatefully partisan. The second experiment demonstrated that the attention economy, by tricking users into believing that their community held more extreme and divisive views than it really did, had the same effect. Showing subjects lots of social media posts from peers that expressed outrage made them more outrage-prone themselves. All it takes is regular scrolls through your anger-filled feed not only to make you feel angrier while you're online, but also to make you an angrier person.

Two other scholars later found that moral-emotional content also leads users to express more calls for violence. They trained a computer to analyze the text of articles and blog posts from across the web, then do the same for user comments posted in response: 300 million comments in all. They found that, across topics or political ideologies, as the number of moral-emotional words in an article increased, commenters grew significantly likelier to threaten or incite violence against some perceived enemy, usually someone named in the article. It was a chilling demonstration of

how portraying people and events in sharply moral-emotional terms brings out audiences' instincts for hatred and violence — which is, after all, exactly what social platforms do, on a billions-strong scale, every minute of every day.

"Online platforms," Brady and Crockett wrote, "are now one of the primary sources of morally relevant stimuli people experience in their daily life." Billions of people's moral compasses potentially tilted toward tribalism and distrust. Whole societies nudged toward conflict, polarization, and unreality — toward something like Trumpism.

Brady did not think that social media was "inherently evil," he told me. But as the platforms evolved, the effects only seemed to worsen. "It's just gotten so toxic," he said. "In college, it was nothing like it is now." It was important for people to remember, he felt, that the designers and engineers, who aim to keep you using their platform for as many minutes and hours per day as possible, "have different goals, I don't want to call them good or bad goals, but goals that might not be compatible with yours."

But for all they had learned, Brady and Crockett were, they knew, only beginning to understand the consequences. What effect did all this distortion, this training, have on our societies and politics, on our species?

Without realizing it, I was stumbling my own way toward an answer. As Brady and Crockett continued to investigate the fun house–mirror distortions of social media psychology throughout 2017, I set out, that fall, for a place much farther away, one that the platforms had expended special effort in ignoring, but that would soon become a byword for their greed, their negligence, and their danger: Myanmar.

Seven

The Germs and the Wind

1. A *Good Deal of Good*

BY THE TIME I landed in Myanmar, the soldiers were already throwing babies into fires. For weeks, the military had waged unrestrained war on the thatched-roof villages that dotted the country's westernmost province. Whole battalions pushed from paddy to paddy as gunships roared overhead. They claimed to hunt insurgents. In reality, they were setting upon a community of one and a half million Muslim farmers and fishermen who called themselves Rohingya.

The soldiers, sent to exterminate the impoverished minority that many of Myanmar's leaders and citizens had come to see as an intolerable enemy within, would arrive at a village, then begin by setting rooftops afire. They lobbed grenades through hut doorways and sent rockets slamming into the walls of longhouses. They fired into the backs of peasants fleeing across the surrounding fields. As the houses burned, the men of the village would be arrayed in a line and shot to death. Families streamed by the hundred thousand toward the border. The soldiers attacked these too. They hid land mines in the refugees' paths. Survivors who made it to relative safety in Bangladesh detailed horror after horror to journalists and aid workers who picked their way through the overcrowded camps.

"People were holding the soldiers' feet, begging for their lives," one woman told my colleague Jeffrey Gettleman. "But they didn't stop, they just kicked them off and killed them." When soldiers came to her village, she said, they demanded she surrender the infant she was cradling. When

she refused, they beat her, ripped her son from her arms, and threw him into an open fire. Then they raped her.

Her story was typical. A twenty-year-old woman told a Human Rights Watch investigator that soldiers had killed her infant daughter in the same way. The soldiers then raped her and her mother. When her sister resisted, they killed her with bayonets. While this was happening, a group of villagers arrived and beat her three teenage brothers to death. Local men often accompanied the soldiers as eager volunteers, swinging hatchets and farm implements. They were Rakhine, the region's other major ethnic group, who, like most in Myanmar, are Buddhist. Their presence hinted at the communal nature of the violence, as well as the groundswell of public pressure that had occasioned it.

Yangon, the historic capital, felt a world away from the killing. It was October 2017, more than three years since I'd last visited what was now a city transformed. Sanctions had been lifted, reward for Myanmar's generals surrendering power to elected lawmakers. Dusty shop stalls had been replaced with air-conditioned shopping malls. Imported cars glided over newly paved streets. Most people had their nose in a smartphone. Middle-class comforts had brought a mood of easy optimism, even pride. But something roiled beneath the surface.

An idealistic young doctor, now his neighborhood's first elected lawmaker, told me that waves of social media misinformation and incitement kept his community constantly on the verge of race riots, or provoked them outright. Days earlier, his constituents, furious over Facebook rumors accusing a local Islamic school of secretly hosting terrorists, had stormed the building as its students sat in class. The kids, terrified, escaped through a back door. And it wasn't just here, a local imam told me in the darkened back room of a friend's house, where he'd insisted on talking for fear of meeting in public. Across the country, madrassas were being forced to close, he said, as similar rumors led to violence or the threat of it. "We are a scapegoat," the imam said.

The head of Myanmar's first real media collective, a jittery reporter back from years in exile, said the country's long-suppressed journalists, finally unfettered, faced a new antagonist. Social media platforms were doing what even the dictatorship's trained propagandists couldn't: producing fake news and nationalist fanfare so engaging, so flattering to readers' biases, that people chose it voluntarily over real journalism. When reporters tried to correct the misinformation flowing online, they became the target of it instead, accused of abetting foreign plots.

Civic leaders told me that social media platforms were pumping the national bloodstream with conspiracies and ultranationalist rage. Citizens who'd marched for an open, inclusive democracy now spent hours posting in groups dedicated to vilifying minorities or to glorifying the country's leaders. The chief of the military, once a reviled symbol of the dictatorship who had stepped down only a few years earlier, now had 1.3 million Facebook fans.

People from all walks of life breathlessly recounted, as unvarnished fact, crazed and hateful conspiracies that they inevitably traced to social media. Buddhist monks insisted the Muslims were plotting to steal Myanmar's water, old ladies that they would not be safe until minorities were purged from their midst, young students that humanitarian groups were arming the Rohingya on behalf of foreign powers. All of them backed the military's campaign — grateful, sometimes gleeful, for the violence being committed on their behalf.

No algorithm could generate hatred this severe out of nothing. The platforms drew on a crisis that had been building since 2012 in the nation's west, where most Rohingya lived. A handful of incidents between Rohingya and Rakhine — a rape, a lynching, a spree of murders — had spiraled into communal riots. Troops intervened, herding civilians who'd been displaced from their homes, mostly Rohingya, into camps. The Rohingya languished. In 2015 thousands attempted to flee, describing growing persecution from neighbors and soldiers alike.

Anti-Rohingya sentiment dated back at least a century, to the early

1900s, when British overlords imported thousands of colonial subjects from the Indian Raj, many of them Muslim. The effort was playbook divide and rule; the newcomers, who filled out the urban merchant class, relied on the British for safety. After the British left, in 1948, independence leaders sought to consolidate their new nation around shared ethnic and religious identity. But Myanmar's diversity made this difficult; they needed an enemy to rally against. Political leaders promoted colonial-era suspicions of Muslims as alien interlopers sponsored by foreign empires. In truth, however, merchant-class Indians imported by the British had mostly fled in 1948 or shortly thereafter, so leaders sublimated national ire to an unrelated group of Muslims: the Rohingya. To sell the ruse, the Rohingya were classified as illegal immigrants, a declaration of state-sponsored hate later reiterated even by Aung San Suu Kyi, the Nobel-winning democracy icon who became Myanmar's first elected leader.

When some Rohingya and Rakhine clashed in 2012, she was still consolidating her hold on politics. She seized on the incident, emphasizing the Rohingya's supposed danger to Myanmar's "real" citizens. But over the next few years, public rage at poor Rohingya farmers soared far beyond what even she had encouraged. By August 2017, when sporadic violence between soldiers and a handful of Rohingya rebels culminated in a midnight insurgent attack on several police posts, much of the country was screaming for blood. A few days later, the military complied, launching their genocide.

How had sentiment, even long simmering, escalated to such extremes? Fearmongering leaders and sectarian clashes were, after all, nothing new here. There was something different at play, something new. Two years earlier, David Madden, the Australian who ran Myanmar's largest tech-startup accelerator, had flown to Facebook's headquarters to give the company's executives an alarm-ringing presentation. By this time, it had been a year since the riots in Mandalay, when the danger should have been unignorable. He detailed rising anti-Muslim incitement on the platform, seemingly unchecked by moderators, however many there were, who were

supposed to scrub dangerous content. He warned that Facebook could soon be used to foment genocide. But there was little indication that Facebook heeded his warning, with hate speech only growing more common. Viral posts, one after another, reported that seemingly innocent Muslim families were really terrorist sleeper cells or foreign spies. "Myanmar will soon be seized by 'Muslim Dogs,'" one read. The posts were shared thousands of times, numbers that would have been hard to achieve in such a small country without an algorithmic boost.

Even Myanmar government officials warned that Facebook-driven hate speech could undermine the country's stability as extremists gained vast new online audiences. By the fall of 2015, Wirathu, the monk once called "the Burmese bin Laden," had 117,000 followers — a small number in the United States, but a large one in a country Myanmar's size and this early in its digital adoption — to whom he pushed steady conspiracy and hate. An ally of Wirathu's, the nationalist politician Nay Myo Wai, ran popular accounts that spread open incitement. He had said of the Rohingya in a speech that year, "I will keep this short and direct. Number one, shoot and kill them. Number two, shoot and kill them. Number three, shoot and bury them."

A Washington DC think tank analyzed a sample of 32,000 Myanmar Facebook accounts, everyday users, finding their pages awash in hate speech and misinformation. One popular meme showed graphic bestiality covered in Arabic script, another of the prophet Mohammad being orally penetrated. Another claimed to show evidence of Rohingya committing cannibalism; the image was in fact taken from a video game marketing stunt. It was shared nearly 40,000 times. Another, falsely claiming that Rohingya were smuggling weapons into Myanmar, was shared 42,700 times. "It's time to kill all kalars," one user wrote, using a slur for Rohingya. Another responded, "We will behead ten thousand kalars' heads." Another: "For the next generation, burn all Muslim villages nearby."

The report was published in early 2016, another voice in a chorus warning Facebook that it was imperiling a society it did not understand.

That June, the company, much as it had in 2013 after brushing off warnings of impending violence that quickly proved accurate, scaled up in Myanmar anyway, launching "Free Basics," which allowed locals to use Facebook's smartphone app without paying data charges. Within months, 38 percent of people in the country said they got most or all of their news via Facebook. As things worsened, six months before the genocide, Madden flew to Facebook's headquarters for a second time. Again he warned that the platform was pushing the country toward mass violence. Nothing appeared to change, even as the killing began.

"I have to thank Facebook because it is giving me the true information in Myanmar," the administrator of a village that had banned Muslims told my colleague Hannah Beech two months into the bloodshed. "Kalar are not welcome here," he said, "because they are violent and they multiply like crazy." Extremist pages espousing these views remained hyperactive throughout the bloodshed. They were a digital update of Radio Milles Collines, which had broadcast calls for genocide in 1990s Rwanda. But this Genocide Radio was built on infrastructure owned by wealthy American tech companies, amplified not by militia-controlled broadcast terminals but by algorithms run out of Silicon Valley.

"There has never been a more powerful tool for the rapid dissemination of hate speech and racist-nationalist vitriol than Facebook and other social media," Ashley Kinseth, a human rights worker in Myanmar, wrote amid the killing. For all the parallels with Radio Milles Collines, she added, "Social media is by all accounts an even faster, more graphic, immersive, 'democratic,' and ultimately dangerous tool for the dissemination of hate speech."

For years after Rwanda's genocide, American officials tormented themselves over hypotheticals. Could American warplanes have destroyed the radio towers in time to stop it? How would they locate the towers amid Rwanda's jungles and mountain passes? How would they secure international authority? In Myanmar, there were never any such doubts. A single engineer could have shuttered the entire network as they finished their

morning coffee. One million terrified Rohingya made safer from death and displacement with a few keystrokes. The warning signs were freely visible. Madden and others had given them the necessary information to act. They simply chose not to, even as entire villages were purged in fire and blood. By March 2018, the head of the United Nations' fact-finding mission said his team had concluded that social networks, especially Facebook, had played a "determining role" in the genocide. The platforms, he said, "substantively contributed" to the hate destroying an entire population.

Three days later, a reporter named Max Read posed a question, on Twitter, to Adam Mosseri, the executive overseeing Facebook's news feed. He asked, referring to Facebook as a whole, "honest question — what's the possible harm in turning it off in myanmar?" Mosseri responded, "There are real issues, but Facebook does a good deal of good — connecting people with friends and family, helping small businesses, surfacing informative content. If we turn it off we lose all that."

The belief that Facebook's benefits to Myanmar, at that moment, exceeded its harms is difficult to understand. Facebook had no Myanmar office from which to appreciate its impact. Few of its employees had ever been. It had rejected the chillingly consistent outside assessments of its platform's behavior. Mosseri's conclusion was, in the most generous interpretation, ideological, rooted in faith. It was also convenient, permitting the company to throw up its hands and declare it ethically impossible to switch off the hate machine. Never mind that leaving the platform up was its own form of intervention, chosen anew every day.

There was another important barrier to acting. It would have meant acknowledging that the platform may have shared some blame. It had taken cigarette companies half a century, and the threat of potentially fatal litigation, to admit that their products caused cancer. How easily would Silicon Valley concede that its products could cause upheaval up to and including genocide?

Myanmar was hardly the first indication of those harms. Though it's easy to forget now, events like the Arab Spring uprisings of 2011 had been,

at the time, viewed as proof of social media's liberating potential. But there were signs of trouble, even then. In 2012, in a bizarre episode in India I'd written about, members of two ethnic groups had, in their mutual fear, spread Facebook and Twitter rumors that the other was planning to attack them. Speculation became certainty, which became misinformation of an imminent attack, which became incitement to strike first. A few inevitably did. Reports of the violence spread widely online, often portrayed, with phony photo proof, as hundreds of times deadlier than it really was. A wave of riots and reprisals, incited on social media, swept across India, pushing 300,000 people into displacement camps. The Indian government blocked access to social platforms and demanded they remove the most dangerous content. When the Obama administration, a longtime Silicon Valley booster, intervened on the companies' behalf, Indian officials relented. The damage had already been done, anyway. Similar violent flare-ups rose in Indonesia. Whole communities glued to Facebook and Twitter. Users rewarded with huge audiences for indulging one another's worst tendencies. A riot, a murder, a village disintegrating into bloodshed, all provoked by xenophobia saturating the platforms.

Eventually, the sunny view of the Arab Spring came to be revised. "This revolution started on Facebook," Wael Ghonim, an Egyptian programmer who'd left his desk at Google to join his country's popular uprising, had said in 2011. "I want to meet Mark Zuckerberg someday and thank him personally." Years later, however, as Egypt collapsed into dictatorship, Ghonim warned, "The same tool that united us to topple dictators eventually tore us apart." The revolution had given way to social and religious distrust, which social networks widened by "amplifying the spread of misinformation, rumors, echo chambers, and hate speech," Ghonim said, rendering society "purely toxic."

By late 2017, as the Myanmar genocide raged on, Chamath Palihapitiya, Facebook's former chief of global growth, speaking at what was expected to be a routine speech to Stanford MBA students, snapped. "I feel tremendous guilt," he said. "I think we all knew in the back of our

minds, even though we all feigned this whole line that there probably weren't any unintended consequences. I think we knew that something bad could happen." Palihapitiya had left Facebook years earlier. But he had helped set the company down the path it remains on today, persuading its chiefs to reengineer both the business and the platform around permanent, globe-spanning growth. The tools they had created to accomplish this were "ripping apart the social fabric," Palihapitiya said. "The short-term, dopamine-driven feedback loops we've created are destroying how society works," creating a world with "no civil discourse, no cooperation; misinformation, mistruth." He urged the would-be engineers and startup founders in the room to take heed. "If you feed the beast, that beast will destroy you," he said. "If you push back on it, we have a chance to control it and rein it in."

This string of breakdowns, their horrifying consistency, including the 2016 U.S. presidential election, suggested more than freak incidents. It hinted at a deeper, perhaps universal transformation wrought by the social networks, of which extreme violence was just a surface-level indicator. I wanted to understand why this was happening, what it revealed about this technology's influence over our world. But a society-wide shift like Myanmar's or America's was driven by too many factors to isolate social media's role. I needed to start with a more self-contained episode, where social media's effects could be isolated, to understand the trend.

I worked with Amanda Taub, a fellow *New York Times* reporter with whom I'd collaborated since 2014, when I'd recruited her to join *Vox*. She'd previously worked as a human rights lawyer, including in Latin America, which made her especially attuned to the warning signs of collective violence. And she shared my fascination with social media, as well as a sense that its influence remained incompletely understood. We put in calls to rights workers, digital monitors, and other trusted contacts. Our question to each of them was whether they had seen unusual upheaval driven by social media. They all had the same answer, whatever continent we reached them on: Yes, more all the time, and why has it taken you all so

long to notice? But collecting information on a long-past incident wouldn't do; memory is imperfect and shaded by bias. Amanda and I needed to see firsthand, to trace back every step and rumor. We asked our contacts to call us if anything combusted outside their windows.

We didn't wait long. In early 2018, someone alerted us to a flash of violence paralyzing Sri Lanka, the teardrop-shaped island nation about the size of Maine off India's southern coast. Whole villages, as if suddenly possessed, had formed into mobs, ransacking and burning their neighbors' homes. The military had been deployed. Though it was unclear what had happened or why, everyone we contacted there named the same culprit: Facebook.

2. *The Tinderbox and the Match*

PAST THE END of a remote mountain road, down a rutted dirt track, in a concrete house without running water but bristling with smartphones, thirteen members of an extended family were glued to Facebook. And they were furious. The fourteenth member of their family had been beaten to death a few weeks earlier. The police said he'd gotten into a traffic dispute that had turned violent. But on Facebook, rumors insisted that his assailants were part of a Muslim conspiracy to wipe out the Sinhalese, Sri Lanka's ethnic majority. The Sinhalese, from the Sanskrit word for "lion," dominate the country's culture and politics. Their lion emblazons its flag. But they had been gripped by a strange racial panic.

"We don't want to look at it because it's so painful," H. M. Lal, the victim's cousin said, his voice trembling. "But in our hearts there is a desire for revenge that has built." When I asked Lal and the rest of his family if they believed the posts were true, all but the elderly, who seemed not to follow, nodded. Did other people on Facebook share their desire for revenge? I asked. Again they nodded. They had shared, and could recite verbatim, memes constructing an alternate reality of nefarious Muslim plots. Though they had not joined in when Facebook groups boasting

thousands of members planned a spree of retaliatory attacks on Muslims, they did not disapprove, either.

"Facebook is important to us because if something is happening somewhere, that's how we find out," one said. "Facebook will tell us about it." Lal, the cousin, agreed. He called Facebook "the embers beneath the ashes" of racial anger that, only days earlier, had brought the country to chaos. "People get provoked into action." This mountain village was our starting point for retracing Sri Lanka's slide into chaos. Facebook, we found, had driven every deadly step. And at every step, as in Myanmar, it had been warned, urgently and explicitly, but refused to act.

We asked the family how it had happened. Everything had "started in Ampara," one said, uttering a name we had seen over and over online. The real Ampara was just another village in a country scattered with them, a few concrete buildings surrounded by open green fields. But the imagined Ampara, constructed from social media rumors, was the epicenter of a plot to destroy the country's Sinhalese.

The Atham-Lebbe brothers knew nothing of the imagined Ampara when, using money they'd saved toiling as manual laborers overseas, they opened a one-room restaurant here. They are Muslim and speak Tamil, a minority language, so they never encountered the Sinhalese-speaking districts of the social web where their town was a symbol of racial peril. So they had no way to anticipate that, on a warm evening in March 2018, the real and imagined Amparas would collide, upending their lives forever.

During that night's dinner rush, a customer began yelling in Sinhalese about something he had found in his beef curry. Farsith, the twenty-eight-year-old brother running the register, ignored him. He didn't speak Sinhalese. And drunk customers, he'd learned, were best ignored. He wasn't aware that, the day before, a viral Facebook rumor had claimed, falsely, that police had seized 23,000 sterilization pills from a Muslim pharmacist here. If he had, Farsith might've understood why, as the customer grew more agitated, a crowd began to form.

The men circled Farsith, slapping his shoulders, yelling a question that Farsith couldn't quite understand. He grasped only that they were asking about a lump of flour in the customer's curry, using the phrase "Did you put?" He worried that saying the wrong thing might turn the crowd violent, but so would saying nothing. "I don't know," Farsith said in broken Sinhalese. "Yes, we put?"

The mob, hearing confirmation, collapsed onto Farsith and beat him. They had been asking if he'd put sterilization pills in the food, as they'd all seen on Facebook. Leaving him bloody on the floor, they pulled down shelves, smashed furniture, ripped appliances from the walls. Dozens of men from the neighborhood, having heard that the Facebook rumors were true, joined in. They marched to the local mosque, which they set on fire while the imam hid in his smoldering office, waiting to die.

In an earlier time, this calamity might have ended in Ampara. But someone in the mob had taken cell-phone video of Farsith's admission: "Yes, we put." Within hours, it was shared to a Sri Lankan Facebook group called the Buddhist Information Center, which had won a fervent following by claiming to provide true information about the Muslim threat. The page published the shaky, eighteen-second clip as proof of the Islamophobic memes it had hosted for months. Then the video spread.

As in Myanmar, social media had been initially received as a force for good in Sri Lanka. It kept families in touch even as many worked abroad to send money home. Activists and elected leaders credited it with helping to usher in democracy. And thanks to zero-rating programs, the same strategy Facebook had used in Myanmar, millions of people could access the services for free.

Zero-rating had grown out of a peculiarity of Silicon Valley economics: the mandate for perpetual user growth. Poorer countries are not particularly lucrative for platforms; advertisers pay little to reach consumers making a few dollars a day. But by spending aggressively now, the companies could preemptively dominate a poor country's media and internet

markets, where they would face few competitors. They could tell investors that revenue was primed to explode in ten or twenty years, as consumers there entered the middle class.

Facebook, WhatsApp, Twitter, Snapchat, and others launched zero-rated services in dozens of countries, from Colombia to Kenya, where they had no footprint and little familiarity, reasoning they would learn as they went. They might contract a few local English teachers to translate essentials like the "Add friend" button. They would outsource the rest to — what else? — machine-learning algorithms. If the translations were wrong, they'd find out by tracking user behavior.

"As the usage expands, it's in every country, it's in places in the world and languages and cultures we don't understand," Chris Cox, Facebook's chief product officer, boasted in 2013. He cited one in particular: Myanmar, where he'd heard that Facebook already dominated locals' access to news. There was, they told themselves, whether out of ideological fervor or financially motivated disinterest, no need to monitor or even consider the consequences, because they could only be positive.

This was more than hubris. It drew on an idea, suffusing the Valley, that had originated with Peter Thiel, Facebook's foundational investor: "zero to one." It was a mandate, commercial and ideological, for companies to invent something so new that there was no market for it — starting at zero — and then control that market absolutely, a field with one entrant. "The history of progress is a history of better monopoly businesses replacing incumbents," Thiel wrote. Intel and processors. Apple and personal computers. Uber and private taxis. Facebook and social networking.

A monopoly, liberated from competition, would be freed to invest in innovation, bettering all mankind, he argued. This was baseless: monopolies, as a rule, leverage their power to deliver less and less value while extracting greater and greater rents from consumers. But it resonated in the Valley, whose citizens reinterpreted the infinite-growth business model, imposed by investors a few years earlier with the rise of cloud computing, into a glorious mission, the continuation of '90s-era internet liber-

ationism. It implied that overrunning whole societies, blindly trampling whatever had come before, was not only acceptable but necessary.

Such an outcome, far from negative, was considered a gift to the world. The tech industry would bring about nothing less than the "next step" in our journey as a species, Zuckerberg wrote in a 6,000-word essay published a year before I arrived in Sri Lanka. In perhaps the last gasp of Valley utopianism, he pledged that Facebook would provide the "social infrastructure" of a new era, elevating us beyond mere "cities or nations" into "a global community." This would enable "spreading prosperity and freedom, promoting peace and understanding, lifting people out of poverty," even "ending terrorism, fighting climate change, and preventing pandemics."

The results on the ground bore little resemblance to these starry-eyed visions. In the days after the Facebook-inspired mob ravaged Ampara, calls for genocide saturated the platform. "Kill all Muslims, don't even save an infant," one post said. There were hundreds like it, all inspired by the video of Farsith saying, "Yes, we put." A Facebook-famous extremist urged his followers to descend on a local Muslim enclave and "reap without leaving an iota behind." Members of a local human rights group, huddled in a small office in the capital city, Colombo, marked down every post, tracing a network of hate. They planned to pass it all along to Facebook. The researchers were doing Facebook's work for them, they knew, and for free. Volunteer janitors for one of the world's wealthiest platforms. But the company ignored them.

"We have given, for the past four years, data-driven examples of hate. We've given them pages of data," Sanjana Hattotuwa, then a researcher with that rights group, Center for Policy Alternatives, told us. "It's pointless to coordinate with Facebook," he huffed, pacing angrily. Hattotuwa, a familiar face at international technology conferences, had managed to make some connections at the company. But no matter how extreme the incitements to violence, no matter how stridently he warned that the platform was going to get somebody killed, the response was the same: "They

say it doesn't contravene anything. They say please get back to us with more information."

Months earlier, before the devastation in Ampara, one of his colleagues, Raisa Wickrematunge, had spoken at a Stanford forum on social media disinformation. During a coffee break, she cornered a Facebook security manager, Jen Weedon, who'd sat on an earlier panel. She warned Weedon that in Sri Lanka, Facebook was letting open calls to violence, forbidden under the company's own policies, run rampant. The conversation ended inconclusively. After the conference, Wickrematunge sent Weedon a follow-up email, offering to flag dangerous hate speech for Facebook to review — free assistance. She never received a response.

In October 2018, Sri Lankan civil leaders gave Facebook's regional office, which oversees South Asia's 400 million users from India, a stark presentation. Hate speech and misinformation were overrunning the platform, seemingly promoted by its algorithms. Violent extremists operated some of its most popular pages. Viral falsehoods were becoming consensus reality for users. Facebook, after all, had displaced local news outlets, just as it had in Myanmar, where villages were still burning. Sri Lanka might be next. Separately, government officials met privately with Facebook's regional chiefs in Colombo. They pleaded with the company to better police the hate speech on their platform. These posts and pages violated the company's own rules. Why wouldn't Facebook act?

Facebook's position was the same in both meetings. It wasn't enough for someone, even a government minister, to flag a post as hate speech. Facebook, to act, had to verify any rule-breaking itself. But the platform outsourced most of this work to I.T. companies, which did not employ enough Sinhalese speakers to keep pace. Facebook representatives made vague promises about staffing up.

The government officials asked if there was someone they could contact directly in case of an explosion of Myanmar-style incitement. No, the company reps told them. If they saw anything dangerous, they should use the on-site form for reporting rule violations. This directive was maddening. That form, designed for everyday users, was the very same widget to which Hattotuwa and his colleagues had already filed months of increasingly alarmed reports, to almost total silence. All as the calls to violence were getting progressively more specific, naming the mosques and neighborhoods to be cleansed.

3. *What Compels Facebook?*

ACROSS COLOMBO, IN the colonial-era offices housing Sri Lanka's government ministries, the country's information chief, Sudarshana Gunawardana, told us that he and other officials "felt a sense of helplessness." Before Facebook, in times of communal tension, he could meet with civic leaders and media heads, urging messages of calm. Now, everything his citizens saw and heard was controlled by engineers, far away in California, whose local representatives would not even return his calls.

As signs of coming violence mounted, officials rushed out statements debunking the most dangerous rumors. Nobody believed them. They had seen the truth with their own eyes right on Facebook. Gunawardana marked post after post using Facebook's reporting widget. A high-ranking official reduced to begging, via Facebook's submission box, for some anonymous moderator to take notice of his country's spiral toward violence. Every single report was ignored. "There needs to be some kind of engagement with countries like Sri Lanka," Gunawardana said. "We're a society, we're not just a market."

As anger over the Ampara video spread, Facebook extremists directed the rage. One of them, Amith Weerasinghe, whose hatemongering had been rewarded with thousands of followers, seized on the traffic dispute in which Muslim youths had beaten a truck driver — the man whose family we'd met. Weerasinghe circulated memes, shared thousands of times, calling it the first blow in a Muslim uprising. As proof, he promoted the fake-news story about Ampara police seizing many thousands of sterilization pills from Muslim pharmacies. To millions of Sri Lankans stewing in

social media unreality, the supposed confession by Farsith, the restaurant owner, looked like confirmation of everything. The race war was here. A few days after Ampara's riot, the truck driver, still in the hospital, died, which caused online outrage to surge, as it often did, into calls for collective action: true Sinhalese should attend the funeral to show solidarity against the Muslim menace. Busloads arrived at Kandy, the city nearest to the truck driver's village. Some fanned into surrounding towns.

To coordinate movements, Facebook users circulated links to private WhatsApp groups. The Facebook-owned messaging app enables rapid-fire communication, akin to group text messaging for hundreds of people at once, with some viral-friendly twists. Users can forward content from one group to another, enabling posts to spread exponentially. A large WhatsApp group can resemble a mishmash of Facebook, Twitter, and YouTube, filled by viral content copied in from all three. WhatsApp sells itself especially on privacy: end-to-end encryption keeps out prying authorities. There are no fact-checkers or moderators.

The digital researchers joined some of the groups. It wasn't difficult; group names were posted on Facebook hate pages, which operated as openly as newspapers. In one viral WhatsApp video, a man dressed as a monk yelled, "The knife at home is no longer to cut jackfruit. So kindly sharpen that knife and go." In another group, a user shared a photo of a dozen makeshift weapons with a list of targets. He marked two mosques with the word "tonight," and another two with the word "tomorrow." The groups filled especially with content from Weerasinghe. Many shared a video he'd posted on Facebook and YouTube that showed him walking the shops of a town called Digana. Too many of them were owned by Muslims, he said, urging Sinhalese to take the town back. The researchers sent it all to Facebook. No response came.

They watched helplessly as hundreds of Sinhalese posted live from the villages and towns whose streets they filled. Residents hung banners with images of lions out their front windows. It was a message: Sinhalese live

here. Everyone knew what was coming. The first Molotov cocktails flew that evening. For three days, mobs ruled the streets. Going house to house, wherever Muslims lived, they smashed through the front doors, ransacked floor to ceiling, then set the homes afire. They burned mosques and Muslim-owned businesses. They beat people in the street.

In Digana, the town where Weerasinghe had walked in his video, one of those homes belonged to the Basith family. They sold slippers from the first floor and lived on the second. Most had fled. But an elder son, Abdul, had stayed behind and was trapped upstairs. "They have broken all the doors in our house," Abdul said in an audio message he sent to his uncle on WhatsApp. "There are flames coming inside." After a few moments, he pleaded, his voice rising, "The house is burning." His family could not reach the house. Police did not retake Digana until the next morning. They found Abdul dead upstairs.

The country's leaders, desperate to stem the violence, blocked all access to social media. It was a lever they had resisted pulling, reluctant to block platforms that some still credited with their country's only recent transition to democracy, and fearful of appearing to reinstate the authoritarian abuses of earlier decades. Two things happened almost immediately. The violence stopped; without Facebook or WhatsApp driving them, the mobs simply went home. And Facebook representatives, after months of ignoring government ministers, finally returned their calls. But not to ask about the violence. They wanted to know why traffic had zeroed out.

A few days later Amanda and I arrived in Digana, where ashes still blew in the streets. The town, in Sri Lanka's interior of rolling emerald hills and nature preserves, sat only thirty minutes from some of the country's most luxurious resorts. Neighbors watched from tea stalls as a man named Fazal welcomed us into his home, just feet from the shell of the building where his brother, Abdul, had died in the fire. Fazal, who works as an imam, used Facebook for everything, the same as everybody else, he said. I asked him about misinformation and hate online, but he didn't

seem to understand. Facebook simply was. I might as well have asked if he blamed the fires on the wind. I didn't want to press a man in mourning. He put out ice cream for us and left for work.

A young neighbor who had joined us in Fazal's house, Jainulabdeen, told us, once our host had gone, "We expected this." Perhaps not wanting to embarrass Fazal, he had waited to speak up. Like the Basith family, Jainulabdeen was Muslim. But Sinhalese neighbors had warned him days in advance. "Most of them knew," he said. "They knew it from Facebook." When I asked about the video of Weerasinghe, the Facebook extremist, walking Digana to call for Muslims' expulsions, Jainulabdeen snorted and shook his head. "We know him," he said. "He's from the area." On Facebook, Weerasinghe wielded the power to shape reality for hundreds of thousands. But here in his hometown, he was just, Jainulabdeen insisted, "a normal person." His father was a carpenter. The families knew each other. Jainulabdeen's relatives had even asked Weerasinghe's family to intervene. The family, seemingly sharing their concerns, promised to talk to him, but nothing had come of it. He loved being on the internet too much.

Once the mobs dissipated, police arrested Weerasinghe for incitement. Facebook finally shuttered his page. But the Ampara video that had inspired so much violence, of the innocent Muslim restaurant-worker Farsith Atham-Lebbe pressured to confirm a nonexistent race war, remained online. The researchers continued submitting pleas for Facebook to remove it, and the company continued refusing, either ignoring their reports or answering that the content broke no rules.

Farsith was in hiding, we learned, at the other end of the country. While I scrounged a ride out in hopes of meeting him, Amanda headed back toward the capital to chase down the details of a meeting she'd heard about from a source. Earlier that day, Facebook's policy director for South Asia, Shivnath Thukral, had flown in to meet with government ministers, the source had revealed. Now that Sri Lanka had pulled the plug, Facebook was finally making a show of listening.

Thukral was conciliatory, an attendee told Amanda. He acknowledged that Facebook had failed to address the incitement and hate speech that it had been warned about again and again. He promised better collaboration. The next day, Thukral held an off-the-record call with civil representatives. He conceded that Facebook did not have enough Sinhalese-speaking moderators to control misinformation and hate. He again pledged that the company would hire more.

After a few weeks had passed, we asked Facebook how many Sinhalese-speaking moderators they'd hired. The company said only that they'd made progress. Skeptical, Amanda scoured employment websites in nearby countries. She found a listing, in India, for work moderating an unnamed platform in Sinhalese. She called the outsourcing firm through a translator, asking if the job was for Facebook. The recruiter said that it was. They had twenty-five Sinhalese openings, every one unfilled since June 2017 — nine long months earlier. Facebook's "progress" had been a lie.

"We are a government that came to power on a mandate of free expression," Harindra Dissanayake, a presidential advisor in Sri Lanka, told Amanda. He used social media himself. It had pained him to shut off access, if only for a few days. At their best, he said, social media platforms "made things more transparent, gave voice to people who did not have voices." But the past months, he said, had destroyed his faith in the technology he'd once credited with bringing his country democracy. "This idea of social media as an open, equal platform is a complete lie," he now believed. "There is no editor, there is the algorithm."

He stressed that Sri Lanka's divisions predated social media. But these platforms, he warned, brought out the very worst in a society, amplifying its extremes in ways that had never before been possible. "We don't completely blame Facebook," Dissanayake said. "The germs are ours, but Facebook is the wind, you know?" His government was considering regulations or fines, he said. But he knew Sri Lanka's power was modest. Only Americans, he believed, had enough leverage to force change. "You, the

United States itself, should fight the algorithm. What compels Facebook, beyond that?"

The next day, I arrived at the opposite end of the country, where a local teacher who claimed to know Farsith guided me to a small settlement some miles from Ampara, to a row of two-room concrete houses. He pointed to the third from the end.

Farsith, waiting inside, had shaved his beard. Not to hide his faith, he said, but because even in this far-flung village, he could hardly make it a block without being recognized. "People would ask me all sorts of questions," he said. Or shout at him, "You're from the video!" He recounted the riot, his confusion and fear, the mob's fury. "I thought that would be my last day," he said. He'd fled the next morning.

Shy, almost childlike, he seemed off somewhere else. As we talked, he twisted a hand in front of his five-year-old niece in half-hearted play. She pulled and prodded at it, trying to bring his gaze up from the floor. Her father, who had run the restaurant, brought us bananas and tea. The brothers had taken out so many loans to build the shop, he said, that they'd been unable to afford insurance. Now everything was gone but the debt.

"We don't know what to do," Farsith's brother said. Maybe they would return to work construction in Saudi Arabia, which was where they'd saved up money for the restaurant, though that would mean leaving their families behind. "We are waiting on God for guidance."

Farsith sighed. "I don't have any intention of staying here," he said.

I asked him several times about social media. Facebook had turned him into a national villain. It had spread a lie that ruined his family, perhaps now splitting them apart. It had nearly killed him. Even now, he lived in fear of another mob incited by the platform.

Despite all that, he refused to abandon the social networks. With long, empty days in hiding, he said, "I have more time and I look at Facebook much more."

I was shocked. Even if he bore no ill will toward the company whose

platform had upended his family's lives, I said, he knew firsthand that he couldn't believe what he saw there.

It wasn't that he had faith that social media was accurate, he said. "But you have to spend time and money to go to the market to get a newspaper. I can just open my phone and get the news instead." He looked up from the floor, shrugging. "Whether it's wrong or right, it's what I read."

I kept in intermittent contact with Farsith. His family slipped into poverty. Threats continued to follow him. Someone from Facebook got in touch — citing the article that Amanda and I had written — to ask him what had happened. Farsith told the person that he was desperate for a way to feed himself. He was willing to work. The call ended and he never heard from Facebook again. After a year, he had saved up enough to travel to Kuwait, where he began working as a day laborer. He is still there.

Eight

Church Bells

1. Status Threat

IT WAS DURING an interview with Gema Santamaría, a scholar of vigilante violence who had researched strange incidents in her native Mexico, that I realized I would spend years trying to understand how this pattern might be playing out, albeit in less obvious ways or to less obvious effects, around the world, maybe even in the United States, where parallels with Trumpism's rise were only coming into view. She was finding in Mexico the same kinds of outbreaks that researchers in other countries around the globe had been documenting. A Cancun suburb that erupted into violence over online misinformation. A village of quiet families who, after starting a Facebook page for community news that became a hotbed of paranoid rumors, tied up a pair of bewildered traveling pollsters, whom they had accused of plotting to harvest the organs of local children, and set them on fire. Then, in another village, the same pattern, from the details of the rumor to the method of killing, this time claiming the lives of two men who were in town to buy fence posts.

"Social media plays the role that the ringing of the church bells used to play in the past," Santamaría said. "That's the way that people know that a lynching is going to happen." The platforms, she explained, reproduced certain age-old mechanisms by which a community worked itself into collective violence. Lynching, when a group follows its moral outrage to the point of hurting or killing someone — the tyranny of cousins at work — is a communal impulse. A public show of what happens to those transgressing the tribe.

"The aim of it is to communicate," Santamaría said of lynching. The false rumors that consistently spread in advance of mass violence, she believed, were the tell that social media had learned to reproduce that age-old process. More than merely triggering preexisting sentiment, social media was creating it. The rumors were hardly random. "They have a logic to them," she said. "They do not target everyone." Rather, the rumors activated a sense of collective peril in groups that were dominant but felt their status was at risk — majorities angry and fearful over change that threatened to erode their position in the hierarchy. Because the impersonal forces of social change are, for most people, no more defeatable than the weather, social media had stepped in to provide a more corporeal, conquerable villain: feminist bloggers, the religious minority next door, refugees. "This finally is something that you have control over," Santamaría said. "You can actually do something about it."

In Myanmar, social media platforms indulged the fears of the long-dominant Buddhist majority who felt, with democracy's arrival, a shift in the status quo that had long privileged them. In India, it was the Hindu majority, on similar grounds. In 2018, BBC reporters in northern Nigeria found the same pattern, the Fulani majority pitted against the Berom minority, all on Facebook. In America, social media had tapped into white backlash against immigration, Black Lives Matter, increased visibility of Muslims, cultural recalibration toward greater tolerance and diversity. The most-shared rumors, Santamaría pointed out, often had to do with reproduction or population. Sri Lanka and sterilization pills. America and a liberal plot to replace white people with refugees.

The defining element across all these rumors was something more specific and dangerous than generalized outrage: a phenomenon called status threat. When members of a dominant social group feel at risk of losing their position, it can spark a ferocious reaction. They grow nostalgic for a past, real or imagined, when they felt secure in their dominance ("Make America Great Again"). They become hyper-attuned for any change that might seem tied to their position: shifting demographics,

evolving social norms, widening minority rights. And they grow obsessed with playing up minorities as dangerous, manifesting stories and rumors to confirm the belief. It's a kind of collective defense mechanism to preserve dominance. It is mostly unconscious, almost animalistic, and therefore easily manipulated, whether by opportunistic leaders or profit-seeking algorithms.

The problem isn't just that social media learned to promote outrage, fear, and tribal conflict, all sentiments that align with status threat. Online, as we post updates visible to hundreds or thousands of people, charged with the group-based emotions that the platforms encourage, "our group identities are more salient" than our individual ones, as William Brady and Molly Crockett wrote in their paper on social media's effects. We don't just become more tribal, we lose our sense of self. It's an environment, they wrote, "ripe for the psychological state of deindividuation."

The shorthand definition of *deindividuation* is "mob mentality," though it is more common than joining a mob. You can deindividuate by sitting in the stands at a sports game or singing along in church, surrendering part of your will to that of the group. The danger comes when these two forces mix: deindividuation, with its power to override individual judgment, and status threat, which can trigger collective aggression on a terrible scale.

I thought back to a conversation with Sanjana Hattotuwa, the impassioned digital researcher who'd tracked online hate in Sri Lanka. "The cancer has grown such that you're looking at ordinary people," he'd said. "It's disturbing. The radicalization is happening at a very young age." Even schoolkids from perfectly nice families, if they were active with social media, got sucked in, their worlds and worldviews defined by the status threat they encountered online. "This is their initiation into communal relations," he said. "And it's hate. It's really, really bad."

Perhaps this pattern, of status threat running rampant online, helped explain why, in 2016, Trump supporters had fallen so much further down the digital rabbit hole than other Americans. If social media were built to

activate majoritarian identity panic, then America's shrinking white majority — and especially the non-college-graduate or working-class whites who tend to hold their racial identity most closely and who became the bulk of the Trump coalition — would be dangerously susceptible to the same pattern I'd seen in Sri Lanka. Status threat and digital deindividuation on a national scale. By 2018, that tribe had, with a handful of exceptions like the rally in Charlottesville, not yet worked itself up to outright mob violence. But I wondered whether this sort of social media influence might be coming out in other forms, priming people for racial violence in less obvious but still consequential ways.

I soon got an answer. Just as Sri Lanka combusted in March 2018, two German social scientists neared completion on a long project examining the subterranean effects of social media on their country. The study hinted at a shocking revelation, suggesting that events like those in Myanmar and Sri Lanka, far from being unique, were playing out in Western democracies, too, just in subtler ways. To understand it, I traveled to a small town near Düsseldorf, where my colleague Amanda Taub would join me a few days later.

2. *Irony Poisoning*

FOR TWO DAYS in June 2018, a few months after our reporting in Sri Lanka, I wandered the cobblestone streets of Altena, asking a question that brought sober, knowing nods. What happened to Dirk Denkhaus?

Altena, like many other towns in Germany's industrial northwest, was declining, locals would explain, a situation that left young people bored and disillusioned. Germany had recently accepted nearly one million refugees from far-off war zones, which most in Altena had supported. But some had found the influx disorienting. That was the context, they would say, to understand why Denkhaus, a young firefighter trainee who had been considered neither dangerous nor political, had tried to burn down a refugee group house while several families slept inside.

But those I stopped, whether old or young, repeatedly cited another factor they called just as important as the others: Facebook. Everyone here had seen social media rumors portraying the refugees as a threat. They'd encountered the vitriol filling local Facebook groups, a jarring contrast to Altena's physical spaces, where people waved warmly to refugee families. Many here suspected — and prosecutors would later argue — that Denkhaus had isolated himself in an online world of racist paranoia that had gradually changed him.

Altena exemplified a long-suspected but, as of 2018, scantily studied phenomenon: that social media platforms make whole communities more prone to racial violence. The town was one of more than three thousand datapoints in a study that claimed to prove it. Karsten Müller and Carlo Schwarz, researchers at the University of Warwick in the UK, had gathered data on every anti-refugee attack in Germany over a two-year span, 3,335 in all. It had been a volatile period, as Europe's refugee crisis had been followed by a rise in far-right politics. The sheer scale presented an opportunity to isolate social media's influence. In each incident in the study, the researchers analyzed the respective local community, using a handful of key variables. Wealth. Demographics. Political allegiance. Number of refugees. History of hate crime.

One thing stuck out. Towns with higher-than-average Facebook use reliably experienced more attacks on refugees. This held true in virtually any sort of community: big or small, affluent or struggling, liberal or conservative. The uptick did not correlate with general web usage; it was particular to Facebook. Their data boiled down to a breathtaking statistic: Wherever per-person Facebook use rose by one standard deviation above the national average, attacks on refugees increased by about 35 percent. Nationwide, they estimated, this effect drove as much as 10 percent of all anti-refugee violence.

Experts whom I asked to review the findings called them credible and rigorous. Still, the study later attracted criticism for methodological flourishes. To gauge town-by-town Facebook usage, for example, the researchers tracked a battery of indicators, one of which was how many users joined the Nutella fan page. They reasoned that Nutella was universally popular and culturally neutral, making it a useful benchmark. Critics called the choice unserious and unsound. The researchers ironed out the issues in a later redraft. My interest, however, was not in proving out the math at the end of their paper, but in using it as a road map for Facebook's influence. It was why I'd come to Altena, where the researchers had found that Facebook usage and anti-refugee sentiment were both unusually high and at rates in line with the paper's projections. Perhaps Denkhaus represented a deeper shift.

When refugees had first arrived here a few years earlier, in 2015, so many locals had volunteered to help that Anette Wesemann, who'd taken over the local refugee-integration center after giving up her home in bustling Hanover for quiet village life, couldn't keep up. She would find Syrian or Afghan families attended by whole entourages of self-appointed life coaches and German tutors. "It was really moving," she said. But when she set up a Facebook page to organize volunteer events, it filled with anti-refugee vitriol of a sort she'd never encountered off-line. Some posts were threatening, mentioning local refugees by name. Over time, their anger proved infectious, dominating the page. When I mentioned the research linking Facebook to anti-refugee violence, she responded, "I would believe it immediately."

Anti-refugee sentiment is among the purest expressions of status threat, combining fear of demographic change with racial tribalism. Even if few locals truly hated refugees, their posts rose over and over, rewarded for their ability to provoke, like the anti-vaccine content that Renée DiResta had found overwhelming parenting groups. As their hate overran local pages, creating, as usual, a false impression of consensus, more seemed to join in.

Dirk Denkhaus turned out to have experienced a microcosm of this process. When I met with Gerhard Pauli, the region's chief prosecutor, who'd overseen the investigation into Denkhaus, he pulled out a binder

containing hundreds of printouts of Facebook and WhatsApp posts the police had pulled from Denkhaus's cell phone. His slide into extremism, Pauli said, had begun as a joke. He and a friend would exchange racist memes, often borrowed from public Facebook groups, to provoke and shock each other.

"They found themselves joking, addressing one another as 'mein Führer' and such," the prosecutor said, shaking his head. Over time, the sentiment became sincere. "There's a very small distance," Pauli said, "between joke and real." Denkhaus crossed that distance in about six months. "He said to his partner one day, 'And now we have to do something,'" Pauli said. That night he and his friend broke into the attic of a refugee group house and set a fire seemingly intended to kill all inside. Fortunately, the fire fizzled. Police arrested both men the next day.

There's a term for the process Pauli described, of online jokes gradually internalized as sincere. It's called irony poisoning. Heavy social media users often call themselves "irony poisoned," a joke on the dulling of the senses that comes from a lifetime engrossed in social media subcultures, where ironic detachment, algorithmic overstimulation, and dare-to-offend humor prevail. In more extreme forms, sustained exposure to objectionable content, spent going down Facebook or YouTube rabbit holes, can lower people's defenses against it. Desensitization makes the ideas seem less taboo or extreme, which in turn makes them easier to adopt.

In court, Denkhaus's lawyer emphasized that his client had, in his offline life, shown no animus toward refugees before that night. While intended to downplay social media's relevance, this observation instead underscored its power. In the real Altena, overwhelmingly tolerant social norms prevailed. But on Facebook, a closed environment with its own moral rules, Denkhaus had drifted unchecked toward extremism.

Pauli believes that Denkhaus represented a trend. The prosecutor said he was "quite sure" that social media had exacerbated Altena's rise in violence. A few months later its mayor was stabbed by a man who said he was outraged by the town's pro-refugee policies. Police, Pauli said, suspected a

social media link. Local Facebook pages had filled with rage toward the mayor just before the attack. Police hadn't bothered to collect evidence of online influence, though, since the attacker had already confessed. And even if Pauli considered the Silicon Valley giant a kind of unwitting accomplice, he knew that the company was beyond any justice that he could bring.

His office spent more and more time tracking incitement on the platforms. He was growing concerned, he said, about rumors that could spin otherwise normal people into violence. Strangely, as in Mexico and Indonesia and seemingly every other country, they often seemed to turn on mysterious threats to children. "We have lots of situations where somebody saw somebody outside the kindergarten," Pauli said, shaking his head. "Within five minutes it's spreading," he said, "and from post to post, it gets worse. It takes two hours and then you have some lynch mob on the street."

3. *Superposters*

TRAUNSTEIN, A MOUNTAINSIDE town near Austria, is, in many ways, Altena's opposite. Its tourist economy is thriving. Its politics lean liberal. Young people are active in the community. But as in Altena, Facebook use and anti-refugee violence are both unusually high here. I arrived, now joined by Amanda, looking for something in particular. By checking local Facebook groups for the most active and visible posters, we found what's known as a superposter, someone who is thought to embody the ways that Facebook can make a community incrementally more extreme. His named was Rolf Wassermann.

Whatever image you have in your head of the basement-dwelling internet addict, Wassermann is the opposite. Middle-aged and tanned, an artist by trade, sporting a salt-and-pepper beard and an all-black suit, he looks like he stepped out of a TV ad for upmarket beer. Though conservative, he is hardly radical. But he is furiously active online, where he fits the

superposter's archetypal profile. He posts streams of rumors, strident opinion columns, and news reports on crimes committed by refugees. Though none I saw crossed into hate speech or fake news, in the aggregate they portrayed Germany as beset by dangerous foreigners.

"On Facebook, it's possible to reach people who are not highly political," he told us over coffee. "You can build people's political views on Facebook." He described what he said was a typical arc for people he met there. They'd start as not particularly political. They'd begin posting frequently, perhaps thanks to a sudden surfeit of free time, on whatever items appeared on their feeds. They'd join Facebook groups, which is where he often met them. Over time, they'd become more stridently political, he said. Just as he had.

He preferred social media to newspapers or TV, he said, because "Facebook is more honest." For example, on Facebook he had learned, he said, that the number of refugees in Germany and the crimes they'd committed were both higher than the media claimed. And he had done his best to amplify this revelation. "The things people say on Facebook are just more true," he said. As if realizing the absurdity of believing such a thing on pure faith, he laughed, adding, "I assume they are, anyway. I'm not God, I don't know."

Hyperactive users like Wassermann tend to be "more opinionated, more extreme, more engaged, more everything," said Andrew Guess, a Princeton University social scientist. It's a different set of traits than those you might associate with the much-studied, much-interviewed class of social media addicts and early adopters like Adam, the 4chan devotee. Superposters are a breed of their own, and one that the platforms have rendered exceptionally influential. When more casual users open social media, often what they see is a world shaped by superposters. Social media attracts people with certain personality tics that make heavy usage unusually gratifying. Their predominance, in turn, distorts the platforms' norms and biases.

And those defining traits and tics of superposters, mapped out in a

series of psychological studies, are broadly negative. One is dogmatism: "relatively unchangeable, unjustified certainty." Dogmatics tend to be narrow-minded, pushy, and loud. Another: grandiose narcissism, defined by feelings of innate superiority and entitlement. Narcissists are consumed by cravings for admiration and belonging, which makes social media's instant feedback and large audiences all but irresistible. That need is deepened by superposters' unusually low self-esteem, which is exacerbated by the platforms themselves. One study concluded simply, "Online political hostility is committed by individuals who are predisposed to be hostile in all contexts." Neurological experiments confirmed this: superposters are drawn toward and feel rewarded by negative social potency, a clinical term for deriving pleasure from deliberately inflicting emotional distress on others. Further, by using social media more, and by being rewarded for this with greater reach, superposters pull the platforms toward these defining tendencies of dogmatism, narcissism, aggrandizement, and cruelty.

In an unintended 2015 test of this, Ellen Pao, still Reddit's chief, tried something unprecedented: rather than promote superusers, Reddit would ban the most toxic of them. Out of tens of millions of users, her team concluded, only about 15,000, all hyperactive, drove much of the hateful content. Expelling them, Pao reasoned, might change Reddit as a whole. She was right, an outside analysis found. With the elimination of this minuscule percentage of users, hate speech overall dropped an astounding 80 percent among those who remained. Millions of people's behavior had shifted overnight. It was a rare success in combating a problem that would only deepen on other, larger platforms, which did not follow Reddit's lead. They had no interest in suppressing their most active users, much less in acknowledging that there might be such a thing as too much time online.

Could superposters alter not just what showed up in people's feeds, but their very sense of right and wrong? I put the question to Betsy Levy Paluck, who had won a MacArthur Foundation "genius grant" for her work exploring how social norms influence behavior. I expected her to cite her research on, say, communal violence in Rwanda. Instead, she wanted

to talk about school bullying. Schoolkids bully or don't, she found in a long investigation, based largely not on whether they expect punishment or think the target deserves it, but on whether it feels moral to them. Either bullying felt permissible, even righteous, or it felt wrong, and that internal barometer was what mattered most. But how does our moral barometer become set? We like to think of ourselves as following an innate moral code, derived from lofty principles, lived experience, the advice of a trusted elder. In truth, studies find over and over, our sense of right or wrong is heavily, if unconsciously, influenced by what we believe our peers think: morality by tribal consensus, guided not by some better angel or higher power but by self-preserving deference to the tyranny of cousins.

In an experiment in rural Mexico, researchers produced an audio soap opera whose story discouraged domestic violence against women. In some areas, people had the soap played for them privately in their homes. In others, it was broadcast on village loudspeakers or at community meetings. Men who listened at home were just as prone to domestic violence as they had been before. But men who listened in group settings became significantly less likely to commit abuse. And not out of perceived pressure. Their internal beliefs had shifted, growing morally opposed to domestic violence and supportive of gender equality. The difference was in seeing their peers absorb the soap opera. The conformity impulse — the same one that had led Facebook's first users to trick themselves into fuming over the news feed — can soak all the way to the moral marrow of your innermost self.

Most of the time, deducing our peers' moral views is not so easy. So we use a shortcut. We pay special attention to a handful of peers whom we consider to be influential, take our cues from them, and assume this will reflect the norms of the group as a whole. The people we pick as moral benchmarks are known as "social referents." In this way, morality is "a sort of perceptual task," Paluck said. "Who in our group is actually popping out to us? Who do we recruit in our memories when we think about what's common, what's desirable?"

To test this, Paluck had her team fan out to fifty-six schools, identifying which students were influential among their peers as well as which students considered bullying to be morally acceptable. Then she picked twenty or thirty students at each school who seemed to fit both conditions: these were, presumably, the students who played the greatest role in instilling pro-bullying social norms in their communities. They were asked to publicly condemn bullying — not forced, just asked. The gentle nudge to this tiny population proved transformative. Psychological benchmarks found that thousands of students became internally opposed to bullying, their moral compasses pulled toward compassion. Bullying-related disciplinary reports dropped by 30 percent.

Social media platforms place us all in a version of Paluck's school experiment. But, online, our social referents, the people artificially pushed into our moral fields of vision, are the superposters. Not because they are persuasive, thoughtful, or important, but because they drive engagement. That was something unique to platforms like Facebook, Paluck said. Anyone who got a lot of time on the feed became influential. "In real life, some people might talk a lot but not be the most listened to. But Facebook," she said, "puts them in front of you every time."

And social media doesn't just surround you with superposters. It displays their messages on vast, public forums, where you know that everyone else sees them, too, like the loudspeakers in Mexican villages that had demonstrated such power to alter a community all at once. In Germany, social media appeared to have elevated a class of superposters like Wassermann who gave sitewide users the impression that social norms were more hostile to refugees and more conspiratorial than they really were. It was Facebook's 2006 "Against News Feed" imbroglio, now elevated to an entire nation's political psyche, and directed at millions of the country's most vulnerable residents. Even if none of those superposters explicitly endorsed violence, Paluck said, the aggregate effect of their anti-refugee, anti-government messaging likely made vigilante violence feel tolerated, even encouraged.

That afternoon, at a Traunstein community event, a teacher named Natascha Wolff perked up when she heard me asking about social media. Wolff taught at a vocational school, she said, with a mix of German- and foreign-born students. In recent months, the German kids had veered, almost uniformly, toward strident anti-refugee hostility she'd never encountered before. There were, she knew, likely many reasons for this. But whenever she asked where they'd learned the phony statistics or hateful claims they repeated to one another with alarm, she got the same answer: Facebook.

Any rumor or tidbit on Facebook disparaging foreigners, she said, "gets around fast. People feel confirmed in their viewpoint." She added, whipping her arm up and down to mimic someone slamming a keyboard, "It's just, 'like, like, like.'" If she challenged a false claim, she always got the same response: "Everybody knows this is true." But often the students were wrong about that, too; many in Traunstein rejected the rumors as false. Wolff worried that this Facebook bubble, the false communal consensus, had consequences. Her refugee students had coffee dumped on them in the street, garbage thrown on them from car windows. Casual, light-of-day violence one only attempts with the assumption that it will be tolerated.

Violence born on social media had grown so common that the police had begun treating the platforms as an ongoing threat to public safety. "Facebook is not just like a pinboard where people hang things and others read them," a local police inspector named Andreas Guske told us over coffee the next day. "Facebook, with its algorithm, influences people." Guske, a veteran detective, slightly graying, began to take social media seriously as a threat in 2015, during a nearby Group of Seven summit. When protesters swept in, he noticed platforms filling with rumors, some of which whipped the crowds into paranoid frenzies. The next year, attacks on refugees seemed to rise in concert with online hate speech. He retooled the team overseeing department communications to fight back, online and off. They thought of themselves as public-health workers, inoculating communities against viral misinformation and its consequences.

In one recent case, Guske told me, Facebook had swirled with claims that a group of Muslim refugees in a town near Traunstein had dragged an eleven-year-old girl to a pedestrian underpass and raped her. The rumor, though false, provoked waves of outrage as Facebook pushed it out across Germany. When police denied the story, users insisted that politicians had ordered them to cover it up. The rumors had begun, Guske's team found, after police arrested an Afghan immigrant accused of groping a seventeen-year-old girl. As Facebook users relayed the incident, some added details that shocked or outraged, which sent those versions rocketing past the truth. One assailant became several. A groping became a rape. A teenage victim became an adolescent.

The police posted statements on Facebook and Twitter debunking the rumor by reconstructing its spread. If the police could show how the platforms distorted reality, Guske believed, people would be persuaded to reject what they'd seen there. But he also knew that, on social media, a sober fact-check would never rise as high as a salacious rumor. So his team identified locals who had shared the rumor early in its spread, then showed up at their homes with evidence that they had gotten it wrong. He urged them to publicly disavow their claims, hoping to turn the platforms' own promotion systems against the misinformation. All but one removed or corrected their posts as he'd requested. But they could never keep up with the platforms, whose poisonous output, he feared, was only accelerating. And he lamented that Facebook, at that point a $500 billion company, left it to overworked police departments to manage the risks they created. "It's hard to prevent fake news, because once Facebook pushes it..." he trailed off, shaking his head. "What more can you do?"

That afternoon, as Amanda interviewed locals across town about social media, I met in a nearby park with a young woman who'd attended Wolff's vocational school. She came with a friend, who brought her toddler. Both women, polite but guarded, described themselves as not very political. Neither read the news except for what they saw on Facebook, which they checked frequently. Once I asked how they felt about refugees,

it was all they wanted to talk about. Refugees were violent, they were rapists, and many sympathized with extremists, they said. They recounted lurid, implausible stories of refugee crimes hidden by the government. They had read all about it on Facebook, which was where they often discussed the "refugee situation," one said.

Traunstein leans liberal but is politically split, and I asked the woman if she ever got into arguments about refugees online. She seemed confused by the question. "Everyone feels this way," she said. Her filter bubble, unanimous in fear, had become her reality. She, like Wassermann and his online friends, like Wolff's other students, like the locals that Guske implored to take down racist falsehoods, were the submerged mass of an iceberg of society-wide social media radicalization. Denkhaus, the firefighter-arsonist, was just its tip. There were countless other Germans who had also grown more xenophobic, more conspiratorial, more nationalistic. Most would never resort to violence. But their collective drift had deeper consequences, pulling invisibly at society's mores and politics. In a wealthy democracy like Germany, the result might not be as obvious as a lynch mob or a riot. It might be worse. The country's political center was collapsing. The German far right was rising.

"One of the students in my school was sent back to Africa," the woman said approvingly. The deportation had been over an error in his immigration paperwork. "They should all be sent back."

4. *Going Dark*

THE GERMAN RESEARCHERS at the University of Warwick knew that one element of their theory — causality — needed special attention. Could it be proven that Facebook usage and anti-refugee violence rose in tandem specifically because the former caused the latter? They hit on the idea of examining every significant internet outage in the period their study covered. German internet infrastructure tends to be localized, making outages common but isolated. Each was an opportunity to test causality: if

depriving a community of Facebook suddenly decreased locals' violence toward refugees, it would suggest that Facebook drove some attacks.

Sure enough, whenever internet access went down in an area with high Facebook use, attacks on refugees dropped significantly. The same drop did not occur, however, when areas with high internet usage but only average Facebook usage suffered an outage, suggesting that the violence-provoking effect was specific to social media, rather than from the internet itself. And violence dropped by the *same rate* — 35 percent — at which the study had suggested Facebook boosted such attacks. The researchers stressed that this was not definitive in itself, just an exercise by which to check their conclusions. But it was a striking indication that they were onto something — and an opportunity to consider, with a rigor that one-off shutdowns like Sri Lanka's could not provide, what happens when social media goes away.

"The world got smaller, a lot changed," Stefania Simonutti said, recalling the outage that had blanketed her Berlin suburb for several days to a few weeks, depending on the block. The suburb, Schmargendorf, feels like a haven from the forces of hate. Diverse, middle-class families stroll boutique-lined avenues and upscale farmers' markets. But Facebook usage is high here. So are anti-refugee attacks — except during the outage.

Simonutti, asked how she'd coped, opened her mouth and pressed her palms to her cheeks in a pantomimed scream. She'd lost touch with family abroad — and with the news, for which she trusted only social media. "Many people lie and fake things in the newspapers," she said. "But with the internet, I can decide for myself what to believe and what not." Forced to forgo the social media conspiracies she liked to follow online, she said, she filled the empty time relaxing with her family.

Everybody seemed to remember the outage. Esperanza Muñoz, a cheery, freckled woman who'd moved here from Colombia in the 1980s, had found it relaxing. She socialized more with neighbors and followed the news less. Her daughter, a medical student, said that she hadn't realized how much anxiety the platforms caused her until she went for a few

days without them. The outage, she said, had driven home the extent to which "when news spreads on Facebook, it's made more provocative." Her mother agreed. When her native Colombia had held elections a few weeks earlier, she said, her news feed, dominated by fellow Colombians, had filled with partisan bickering and outrage — and, as if by some script, with fearmongering about refugees.

Earlier that year, in April, Zuckerberg had given an interview to *Vox*'s editor-in-chief, Ezra Klein, who pressed him on the genocide in Myanmar. As evidence for Facebook's progress, Zuckerberg said that, at the height of the bloodshed, the company's security team had identified users in Myanmar inciting violence on Facebook Messenger. "Now, in that case, our systems detect that that's going on. We stop those messages from going through," he said. "But this is certainly something that we're paying a lot of attention to."

After the interview was published, Myanmar rights groups replied with a furious open letter. In fact, they said, they were the ones — and not Facebook — who had found the chain-letter-style messages fomenting violence. And because they lacked Facebook's internal tools for automatically monitoring the platforms, they had been able to ferret them out only through what they stressed was the cumbersome and woefully insufficient method of a manual hunt. Even then, the rights groups had still been forced to barrage Facebook with days of warnings before someone in the company finally acted. But it was too late. The users, apparently acting on these viral messages, had already organized and executed three separate attacks, one of which involved attempting to burn down a school. The episode, the groups said, underscored Facebook's "overreliance on third parties, a lack of a proper mechanism for emergency escalation, a reticence to engage local stakeholders around systemic solutions, and a lack of transparency." Zuckerberg sent the groups an email apologizing, though only for failing to credit them by name, which, the rights workers emphasized in their response, had not been their primary concern.

That August, the United Nations issued its formal report on the genocide. It called the role of social media, particularly Facebook, "significant." But Facebook still refused to share its data with UN investigators, the investigators said, impeding their ability to understand how the genocide had happened and, therefore, how to prevent another. "You can't just snap your fingers and solve these problems," Zuckerberg said a month later. "It takes time to hire the people and train them, and to build the systems that can flag stuff for them." But of course, in both Myanmar and Sri Lanka, Facebook had met warnings of impending violence not with any flurry of new safeguards or moderator hires but with months of inaction. Now, again, nothing appeared to change, a Myanmar-based digital-monitoring group told me. Facebook had solicited the group to monitor for rising online incitement or other dangers. But the company mostly ignored the group's reports, no matter how urgent. Facebook, they believed, had hired them as an empty PR sop.

Adam Mosseri, the executive who had overseen the all-powerful news feed during the Myanmar and Sri Lanka killings, was promoted to vice president of Instagram, then its president. Jen Weedon, the Facebook security-policy manager who had not answered the Sri Lankan researcher's warnings of the coming bloodshed, was promoted as well. Earnings exceeded a record $55 billion that year, up nearly 40 percent from the year before.

"The business model is what got us into trouble," Hany Farid, a UC Berkeley computer scientist who had consulted with governments and rights groups on emerging dangers on the social web, told me later that year. "Four hundred hours of YouTube uploaded every minute. A billion uploads to Facebook a day. Three hundred million tweets a day. And it's sort of a mess," he said. "The tech companies, I wouldn't even say they fell asleep on the job. I'd say they had their eyes wide open. I think they knew exactly what they were doing. They knew the poison was on the network. They knew they had a problem. But it was all about aggressive growth. That's where the problem started from."

Farid took a breath, returning to the topic that I'd called about, a

specialized technology that the platforms used. But later, near the end of a technical explanation, as he stumbled into a reference to YouTube, his voice rose again. "YouTube is the worst," he said. Of what he considered the four leading web companies — Google/YouTube, Facebook, Twitter, and Microsoft — the best at managing what he'd called "the poison" was, he believed, Microsoft. "And it makes sense, right? It's not a social media company," he said. "But YouTube is the worst on these issues," he repeated.

It had been a year of scandal and controversy around Facebook, widely taken as the most influential platform. But Farid's admonition resonated because, even as I investigated the effects of Facebook in Sri Lanka and Germany, I had been hearing the same from digital experts, rights groups, and others: look at YouTube. "YouTube is the most overlooked story of 2016," Zeynep Tufekci, a University of North Carolina sociologist tweeted a year after the election. "Its search and recommender algorithms are misinformation engines." She later called it "one of the most powerful radicalizing instruments of the twenty-first century." Danah Boyd, the founder of a tech-focused think tank, agreed, telling my colleague Amanda, "YouTube is perhaps the most troubling platform we have out there right now."

More and more, stories about strange, destabilizing occurrences — a rising hate group, a dangerous new medical rumor, a lonely kid turned shooter — were mentioning YouTube. I'd barely finished transcribing my notes from Germany when, a few weeks after my conversation with Farid, something happened there that made immediately clear why he'd issued his warning.

Nine

The Rabbit Hole

1. *The YouTube Riot*

NEO-NAZIS HAD HELD the streets of his town on and off for two days when Sören Uhle, a trim and bespectacled municipal official, began to get strange phone calls from reporters. As far as Uhle knew, two Middle Eastern refugees had stabbed a local man during an argument, killing him, which far-right groups had seized on to encourage people to flood into his city. Now, reporters were telling him, it turned out that the refugees had actually killed not one but two men. They had also been molesting a local woman; their victims had died trying to protect her. Could Uhle comment? And could he also explain why politicians were secretly paying locals to attend an upcoming counter-protest?

Uhle was dumbstruck. The revelations were all false. "This was new," he said. "It's never happened to me before that mainstream media, big German newspapers and television channels, ask me about false news and propaganda that had clearly become so pervasive that people just bought it." It was August 2018. The mobs that were swarming Chemnitz, his city of a quarter-million people in eastern Germany, had been organized on social media, he knew. Maybe the misinformation had been, too.

In Berlin, just up the autobahn, a digital researcher named Ray Serrato was arriving at the same conclusion. Like everyone in Germany, he'd been glued to reports from the riots — an out-of-nowhere show of neo-Nazi strength so dramatic that Chancellor Angela Merkel had condemned them. Then his wife's uncle showed him a strange YouTube video. Two middle-aged men, one in dreadlocks and a black beanie, told the camera

that the rioters were not neo-Nazis at all, but Muslim refugees. The video, posted by an obscure fringe group, was rambling and cheaply produced. Yet it had nearly half a million views — far more than any news video on the riots. How was that possible?

Curious, Serrato applied a set of techniques he used in his day job, tracking online hate speech in Myanmar for a democracy-monitoring group. He started with a dozen recent videos on Chemnitz, then, on each, scraped YouTube's recommendations for what to watch next. Then he did the same for those videos, and so on. It revealed a network of about 650 videos: the YouTube-cultivated ecosystem of Chemnitz content. Disturbingly, YouTube's recommendations clustered tightly around a handful of conspiracy or far-right videos. This suggested that any user who entered the network of Chemnitz videos — say, by searching for news updates or watching a clip sent to them by a friend — would be pulled by YouTube's algorithm toward extremist content. Asked how many steps it would take, on average, for a YouTube viewer who pulled up a Chemnitz news clip to find themselves watching far-right propaganda, Serrato answered, "Only two." He added, "By the second, you're quite knee-deep in the alt right."

Recommendations rarely led users back to mainstream news coverage, or to liberal or apolitical content of any kind. Once among extremists, the algorithm tended to stay there, as if that had been the destination all along. It even led from Chemnitz videos to unrelated far-right topics — white nationalism, anti-Semitic conspiracies — much as Facebook had steered Renée DiResta from anti-vaccine pages to entirely separate fringe causes. One typical video called Trump a pawn of the Rothschild banking family. Though Serrano considered the videos abhorrent and dangerous, he admitted that something about them was hard to turn off. "That's YouTube's goal," he said. "I stay engaged, ads play. And it works."

This effect, I realized, working with Katrin Bennhold, Berlin bureau chief for the *New York Times*, had helped produce the chaos in Chemnitz. Shortly after the stabbing there, a handful of obscure, far-right YouTubers had posted videos about the incident. One, a blogger named Oliver Flesch,

had only 20,000 subscribers. He did little outreach or promotion beyond his ideological bubble. Yet his videos on Chemnitz accrued hundreds of thousands of views, thanks to heavy promotion by YouTube's recommendation engine.

Serrato found that viewers who watched anything about Chemnitz on YouTube, like a news clip, were quickly recommended into Flesch's channel. Flesch posted fourteen videos on the topic, all of which showed up in YouTube's recommendations, seeding the platform with the very race-baiting falsehoods that Sören Uhle would later be asked about. Other far-right and conspiracy channels quickly picked up Flesch's version of events, turning an isolated street fight into a tale of imperiled white virtue. YouTube's algorithm boosted these, too.

Even Germans who searched Google for news on Chemnitz were directed to YouTube conspiracists. Google often promotes YouTube videos near the top of search results, an act of corporate synergy designed to boost revenue. This means that YouTube's practices don't stay on YouTube; since Google dominates internet searches, these practices influence how virtually anyone on the web finds and accesses news and information.

As YouTube and Google diverted more Germans to videos about Chemnitz rife with falsehoods, interest in the town grew, including among many outside the far right. The YouTube voices getting all this attention called on their rapidly growing followership to show their support for the stabbing victim by going to Chemnitz. Locals said that in the days before the violence, conspiracy theories grew strangely common, whispered at pubs and watercoolers. Then the crowds arrived, frothing to take back the city from foreigners. Soon they rioted, ransacking shops and brawling with police. Many of the rioters credited YouTube with putting them there.

It was Sri Lanka's meltdown, beat-for-beat, in the heart of Europe. But there was one important difference. Social media had, in Sri Lanka, radicalized a real-world social group with a strongly held identity, the Sinhalese. In Germany, however, Chemnitz's rioters were something new. There were certainly hardened neo-Nazis in the crowd, but many

143 follow a bot that retweeted voices: "Exposure to Opposing Views on Social Media Can Increase Political Polarization," Christopher A. Bail et al., *Proceedings of the National Academy of Sciences* 115, no. 37, September 11, 2018.

143 perceive out-groups as monoliths: Social scientists call this the "outgroup homogeneity effect." See, for example: "Out-Group Homogeneity Effects in Natural and Minimal Groups," Thomas M. Ostrom and Constantine Sedikides, *Psychological Bulletin* 112, no. 3, 1992.

144 "false polarization": For a comprehensive account of false polarization, see "The Great and Widening Divide: Political False Polarization and Its Consequences," Victoria Parker, master's thesis, Wilfrid Laurier University, 2018.

144 false polarization is worsening: "On Trolls and Polls: How Social Media Extremists and Dissenters Exacerbate and Mitigate Political False Polarization," presentation by Victoria Parker, Wilfrid Laurier University, 2019.

144 leads people to develop more extreme views: "Thinking Fast and Furious: Emotional Intensity and Opinion Polarization in Online Media," David Asker and Elias Dinas, *Public Opinion Quarterly* 83, no. 3, fall 2019.

144 groups heighten their sensitivity: "The Spreading of Misinformation Online," Michela Del Vicario et al., *Proceedings of the National Academy of Sciences* 113, no. 3, January 19, 2016.

144 "When we encounter opposing views": "How Social Media Took Us from Tahrir Square to Donald Trump," Zeynep Tufekci, *MIT Technology Review*, August 14, 2018.

145 "the problem with Facebook": "Interview with Siva Vaidhyanathan," David Greene, National Public Radio, *Morning Edition*, December 26, 2017.

145 "I would read something": "Screaming into the Void: How Outrage Is Hijacking Our Culture, and Our Minds," National Public Radio, *Hidden Brain*, October 7, 2019.

145 "It was like coming out of a trance": Ibid.

145 a short but influential paper: "Moral Outrage in the Digital Age," Molly J. Crockett, *Nature Human Behaviour* 1, 2017.

147 "I think they need to be extremely": "Mark Warner to Facebook: Tell Me What You Know," Elaine Godfrey, *The Atlantic*, September 28, 2017.

147–148 "how you would try to wrangle": "The Facebook Dilemma," *PBS Frontline*, October 29, 2018.

148 They couldn't help wondering: Ibid.

148 later termed this "ampliganda": "It's Not Misinformation. It's Amplified Propaganda," Renee DiResta, *The Atlantic*, October 2021.

150 according to an MIT Media Lab analysis: "Who's Influencing Election 2016?," William Powers, Medium.com, February 23, 2016.

150 begun posting 200-plus times per day: For an account of Mackey's story, including details from the federal indictment issued against him, see "Trump's Most Influential White Nationalist Troll Is a Middlebury Grad Who Lives in Manhattan," Luke O'Brien, *HuffPost*, April 5, 2018; and "FBI Arrests Prolific Racist Twitter Troll 'Ricky Vaughn' For 2016 Election Interference," Luke O'Brien, *HuffPost*, January 27, 2021.

151 shared 36,000 times: "Debunking 5 Viral Images of the Migrant Caravan," Kevin Roose, *New York Times*, October 24, 2018.

151 Republicans were shown a false headline: "Shifting Attention to Accuracy Can Reduce Misinformation Online," Gordon Pennycook et al., *Nature* 592, 2021.

151 "Most people do not want to spread": Ibid.

152 heavily favored the candidates: "Does YouTube's Algorithm Promote Populist Candidates in the French Presidential Elections?," Guillaume Chaslot et al., Mediashift.org, April 21, 2017.

153 "the methodology, data and, most importantly": "How an Ex-YouTube Insider Investigated Its Secret Algorithm," Paul Lewis and Erin McCormick, *The Guardian*, February 2, 2018.

153 "Our only conclusion is that": Ibid.

154 reshaping not just online behavior: Except where otherwise noted, all subsequent references to Brady and Crockett's study in this chapter draw from "The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online," William J. Brady, Molly J. Crockett, and Jay J. Van Bavel, *Perspectives on Psychological Science* 15, no. 4, June 2020.

155 showed participants a fake social media stream: "Attentional Capture Helps Explain Why Moral and Emotional Content Go Viral," William J. Brady, Ana P. Gantman, and Jay J. Van Bavel, *Journal of Experimental Psychology* 149, no. 4, 2020.

156 leads users to express more calls: "Moral-Emotional Content and Patterns of Violent Expression and Hate Speech in Online User Comment," Jeffrey Javed and Blake Miller, working paper, April 2019. (Javed subsequently took a job at Facebook, on a team that optimizes ad placement.)

Chapter 7: The Germs and the Wind

158 would arrive at a village: See, for example: *Massacre by the River: Burmese Army Crimes Against Humanity in Tula Toli*, Human Rights Watch report, December 19, 2017.

158 "People were holding the soldiers' feet": "Rohingya Recount Atrocities: 'They Threw My Baby into a Fire,'" Jeffrey Gettleman, *New York Times*, October 11, 2017.

159 A twenty-year-old woman told a Human Rights Watch: *Sexual Violence Against Rohingya Women and Girls in Burma*, Human Rights Watch report, November 16, 2017.

161 had flown to Facebook's headquarters: "How Facebook's Rise Fueled Chaos and Confusion in Myanmar," Timothy McLaughlin, *Wired*, July 2018.

162 with hate speech only growing more common: All examples in this and the next paragraph from: *Hate Speech Narratives and Facilitators in Myanmar*, Center for Advanced Defense Studies (C4ADS) report, February 2016.

162 analyzed a sample of 32,000: Ibid.

163 38 percent of people in the country: *Survey of Burma/Myanmar Public Opinion*, Center for Insights in Survey Research, April 1, 2017.

163 Madden flew to Facebook's headquarters: McLaughlin, 2018.

163 "I have to thank Facebook": "Across Myanmar, Denial of Ethnic Cleansing and Loathing of Rohingya," Hanna Beech, *New York Times*, October 24, 2017.

163 "There has never been a more": "Genocide in the Modern Era: Social Media and the Proliferation of Hate Speech in Myanmar," Ashley Kinseth, *Tea Circle Oxford*, May 2018.

164 "honest question — what's": Tweet by Max Read (@max_read), March 15, 2018 (since deleted).

164 "There are real issues": Tweet by Adam Mosseri (@mosseri), March 15, 2018 (since deleted).

165 in a bizarre episode in India: "When Is Government Web Censorship Justified? An Indian Horror Story," Max Fisher, *The Atlantic*, August 22, 2012.

165 pushing 300,000 people: "Panic Seizes India as a Region's Strife Radiates," Jim Yardley, *New York Times*, August 17, 2012.

165 rose in Indonesia: See, for example: "Beredar Hoax Penculikan Anak, Gelandangan Disiksa Nyaris Tewas," Fajar Eko Nugroho, *Liputan6*, March 7, 2017. "Justice by Numbers," Sana

Jaffrey, *New Mandala*, January 12, 2017. "The Muslim Cyber Army: What Is It and What Does It Want?" Damar Juniarto, Indonesiaatmelbourne.unimelb.edu.au, 2017.

165 "This revolution started": "Social Media Sparked, Accelerated Egypt's Revolutionary Fire," Sam Gustin, *Wired*, February 11, 2011.

165 "The same tool": "Let's Design Social Media That Drives Real Change," Wael Ghonim, TED Talk, January 14, 2016.

165 "I feel tremendous guilt": "Former Facebook Exec Says Social Media Is Ripping Apart Society," James Vincent, *The Verge*, December 11, 2017.

170 launched zero-rated services: *Free Internet and the Costs to Media Pluralism: The Hazards of Zero-Rating the News*, Daniel O'Maley and Amba Kak, CIMA digital report, November 8, 2018.

170 "As the usage expands": *Facebook: The Inside Story*, Steven Levy, 2020: 435.

170 "The history of progress": *Zero to One*, Thiel and Masters, 2014: 32.

171 a 6,000-word essay: "Building Global Community," Mark Zuckerberg, Facebook.com, February 16, 2017.

Chapter 8: Church Bells

180 She was finding in Mexico: "La Otra Violencia: El Linchamiento de José Abraham y Rey David," Gema Santamaría, *Nexos*, October 22, 2015.

180 Cancun suburb: "Un Ruso sobrevive a un intento de linchamiento en Cancún por insultar a los mexicanos," L.P.B., *El País*, May 20, 2017.

180 village of quiet families: "In Frightened Mexico Town, a Mob Kills 2 Young Pollsters," Alberto Arce, Associated Press, October 22, 2015.

180 in another village, the same pattern: "When Fake News Kills: Lynchings in Mexico Are Linked to Viral Child-Kidnap Rumors," Patrick J. McDonnel and Cecilia Sanchez, *Los Angeles Times*, September 21, 2018.

180 is a communal impulse: For more, see *In the Vortex of Violence: Lynching, Extralegal Justice, and the State in Post-Revolutionary Mexico*, Gema Kloppe-Santamaría, 2020.

181 BBC reporters in northern Nigeria: "Like. Share. Kill," Yemisi Adegoke, *BBC Africa Eye*, November 12, 2018.

181 feel at risk of losing their position: For an explication of the research status threat and its relevance to the Trump coalition, see, for example: "Trump-ing Foreign Affairs: Status Threat and Foreign Policy Preferences on the Right," Rachel Marie Blum and Cristopher Sebastian Parker, *Perspectives on Politics* 17, no. 3, August 2019.

182 "our group identities are more salient": "The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online," William J. Brady, Molly J. Crockett, and Jay J. Van Bavel, *Perspectives on Psychological Science* 15, no. 4, June 2020.

183 tried to burn down a refugee group house: "Eine rechtsradikale Einstellung besteht aus mehr als Fremdenhass," *Der Spiegel*, October 12, 2015.

184 gathered data on every anti-refugee attack: "Fanning the Flames of Hate: Social Media and Hate Crime," Karsten Müller and Carlo Schwarz, *Journal of the European Economic Association* 19, no. 4, August 2021.

186 "irony poisoned": See, for example: "How the Parkland Teens Give Us a Glimpse of a Post-Irony Internet," Miles Klee, *Mel Magazine*, February 28, 2018.

186 Denkhaus's lawyer emphasized: "Brandstifterprozess Altena," Akantifahagen.blogsport.eu, May 31, 2016.

189 "relatively unchangeable, unjustified certainty": "Political Tolerance, Dogmatism, and Social Media Uses and Gratifications," Chamil Rathnayake and Jenifer Sunrise Winter, *Policy & Internet* 9, no. 4, 2017.

189 Another: grandiose narcissism: "Why Narcissists Are at Risk for Developing Facebook Addiction: The Need to Be Admired and the Need to Belong," Silvia Casale and Giulia Fioravanti, *Addictive Behaviors* 76, January 2018.

189 Unusually low self-esteem: "The Relationship Between Addictive Use of Social Media, Narcissism, and Self-Esteem: Findings from a Large National Survey," Cecilie Schou Andreassen, Ståle Pallesen, and Mark D. Griffiths, *Addictive Behaviors* 64, January 2017.

189 "Online political hostility is committed": "The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis," Alexander Bor and Michael Bang Peterson, *American Political Science Review*, 2021.

189 Neurological experiments confirmed: "Snapchat vs. Facebook: Differences in Problematic Use, Behavior Change Attempts, and Trait Social Reward Preferences," Dar Meshi, Ofir Turel, and Dan Henley, *Addictive Behaviors* Report 12, December 2020.

189 She was right, an outside: "The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech," Eshwar Chandrasekharan et al., *Proceedings of the ACM on Human-Computer Interaction* 1, November 2017.

189 exploring how social norms influence: For an accessible overview of Paluck's work, see "Romeo & Juliet in Rwanda: How a Soap Opera Sought to Change a Nation," NPR, *Hidden Brain*, July 13, 2020.

190 Schoolkids bully or don't: "Changing Climates of Conflict: A Social Network Experiment in 56 Schools," Elizabeth Levy Paluck, Hana Shepher, and Peter M. Aronow, *Proceedings of the National Academy of Sciences* 113, no. 3, January 19, 2016.

190 experiment in rural Mexico: "How Does Media Influence Social Norms? Experimental Evidence on the Role of Common Knowledge," Eric Arias, *Political Science Research and Methods* 7, no. 3, July 2019.

191 identifying which students were influential: Paluck et al., 2016.

196 "Now, in that case": "Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next," Ezra Klein, *Vox*, April 2, 2018.

196 "overreliance on third parties": "Open Letter to Mark Zuckerberg," Phandeeyar et al., April 5, 2018.

196 sent the groups an email apologizing: "Zuckerberg Was Called Out Over Myanmar Violence. Here's His Apology," Kevin Roose and Paul Mozur, *New York Times*, April 9, 2018.

196–197 formal report on the genocide: *Report of Independent International Fact-Finding Mission on Myanmar*, United Nations Human Rights Council, August 27, 2018.

197 "You can't just snap": "Can Mark Zuckerberg Fix Facebook Before It Breaks Democracy?" Evan Osnos, *The New Yorker*, September 17, 2018.

197 $55 billion: "Facebook Reports Fourth Quarter and Full Year 2018 Results," press release, Facebook Investor Relations, January 30, 2019.

Chapter 9: The Rabbit Hole

199 "This was new": "As Germans Seek News, YouTube Delivers Far-Right Tirades," Max Fisher and Katrin Bennhold, *New York Times*, September 7, 2018.

200 Serrato applied a set: "Revealed: Facebook hate speech exploded in Myanmar during Rohingya crisis," Libby Hogan and Michael Safi, *The Guardian*, April 2, 2018.