Technology

INSTAGRA

The social-media giant is plagued by bullying. Inside the new plan to fight it

BY KATY STEINMETZ

ETHAN COHEN TRIED TO LAUGH OFF HIS FIRST EXPERIENCE WITH bullying on Instagram. Like many kids his age, the Raleigh, N.C., teen eagerly joined the platform in middle school, and one day he discovered fellow students snickering at an account. Someone—he still does not know who—had started taking surreptitious photos of him and posting them under the username ethan_cohens_neck_vein. The feed was dedicated to jeers about what appeared to be a prominent muscle in his neck. One post compared it to the Great Wall of China. To friends, he dismissed it as a dumb prank, but privately he was distressed. Someone was tailing him and posting mocking pictures for all to see. "The anonymity of it was freaky," says Cohen, now 18. He reported the account multiple times to Instagram. Nothing happened, even though guidelines that govern user behavior forbid mocking someone's physical appearance.

Today, Instagram says, the outcome would be different. More sophisticated reporting tools and moderators would quickly shut the account down. And, in the near future, the company aspires to something far more ambitious: sparing users like Cohen from having to report bullying in the first place by using artificial intelligence to root out behavior like insults, shaming and disrespect. At a time when social-media platforms

CHALLENG

PHOTO-ILLUSTRATION BY DELCAN & COMPANY FOR TIME

M’S

E

are being blamed for a great many problems, and are under pressure to demonstrate they can police themselves, Instagram has declared war on bullying. "We are in a pivotal moment," says its head, Adam Mosseri. "We want to lead the industry in this fight."

It's a logical step. As young people have become glued to the app, bullying has become to Instagram what "fake news" is to Facebook and trolling is to Twitter: a seemingly unstoppable ill that users endure in order to be where everyone else is. By one estimate, nearly 80% of teens are on Instagram and about half of those users have been bullied on the platform. And it gets far worse than neck taunts. In high school, Cohen came out as gay on Instagram and was pummeled by direct messages from a popular student calling him a "faggot" and "failed abortion." Users suffer haunting humiliations and threats of violence. More broadly, bullying on sites like Instagram has been linked to self-destructive behavior.

Sheri Bauman, a counseling professor at the University of Arizona who has spent years studying bullying's causes and effects, calls Instagram a "one-stop shop for the bully" because everything they need is there: an audience, anonymity, an emphasis on appearances and channels that range from public feeds to behind-the-back group chats. Instagram executives acknowledge that as they try to attract more users and attention to the platform, each new feature brings with it a fresh opportunity for abuse. "Teens are exceptionally creative," says Instagram head of public policy Karina Newton.

Mosseri is new to the top job. After Instagram's founders abruptly departed late last year—reportedly amid tensions with parent company Facebook—the longtime Facebook employee took over, having honed his crisis-management skills by overseeing Facebook's News Feed. The 36-year-old aims to define a new era for Instagram, calling the well-being of users his top priority. Tackling bullying gives shape to that agenda. Mosseri is dedicating engineers and designers to the cause. His team is doing extensive research, rolling out new features and changing company protocol, all with bullying in mind. But it's a fight with tangled front lines and plenty of possibilities for collateral damage. Go after bullying too aggressively, and risk alienating users at a time when the company is a bright spot of growth for Facebook. Don't do enough, and risk accusations of placing profits over the protection of kids.

Then there's the technical Everest to climb. Creating artificial intelligence to combat bullying means teaching machines to master an evolving problem with complex nuances. Instagram must also be wary of free-speech issues as engineers create tools optimized to find things they should, without suppressing things they shouldn't. "I do worry that if we're not careful, we might overstep," Mosseri says. But he insists nothing trumps the need to keep the platform civil. "We will make decisions that mean people use Instagram less," he tells TIME, "if it keeps people more safe."

80%

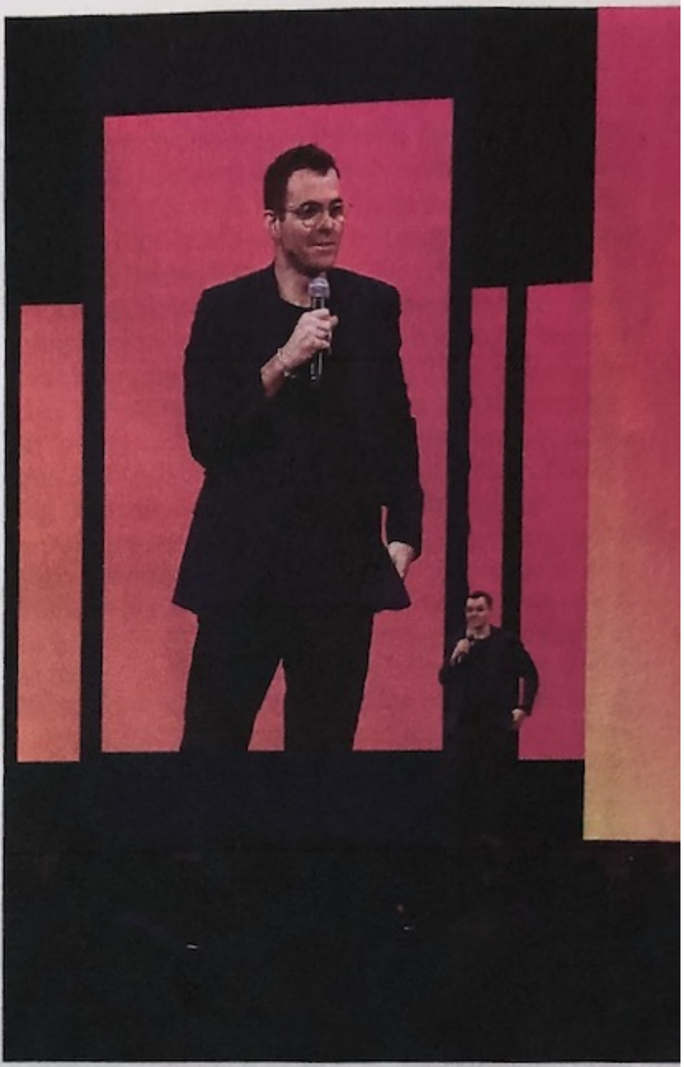
By one estimate, nearly 80% of teenagers are on Instagram, and half of those users have been bullied

10 YEARS

Adam Mosseri became Instagram head after nearly a decade at parent company Facebook

1 BILLION

In 2018, Instagram celebrated the milestone of reaching more than 1 billion users



Facebook stands to profit from every hour people spend on Instagram. If those who associate additional safety measures with constriction go elsewhere, potential revenue leaves with them. When asked if it's in the company's financial interest to take on bullying, Mosseri responds that if Instagram fails to curb it, he will not only be failing users but also failing the business. "It could hurt our reputation and our brand over time. It could make our partnership relationships more difficult. There are all sorts of ways it could strain us," he says. "If you're not addressing issues on your platform, I have to believe it's going to come around and have a real cost."

INSTAGRAM WAS LAUNCHED in 2010 by Kevin Systrom and Mike Krieger, two 20-somethings who hoped users would download a photo-sharing app that made their lives look beautiful. Users did. Within a year, more than 500,000 were signing up each week. But it quickly became clear that the masses were going to use the app for ugly reasons too, and the duo spent time in the early days personally deleting nasty comments and banning trolls. By the time they left, the platform had more than 1 billion users, far too many for humans to monitor. So, like other

Mosseri at Facebook's F8 developer conference in April

JUSTIN SULLIVAN—GETTY IMAGES

social-media platforms trying to ferret out material from terrorist propaganda to child pornography, Instagram turned to machines.

In a quest to make Instagram a kinder, gentler place, the founders had borrowed from Facebook an AI tool known as DeepText, which was designed to understand and interpret the language people were using on the platform. Instagram engineers first used the tool in 2016 to seek out spam. The next year, they trained it to find and block offensive comments, including racial slurs. By mid-2018, they were using it to find bullying in comments too. A week after Mosseri took over in October, Instagram announced it wouldn't use AI just to search for bullying in remarks tacked below posts; it would also start using machines to spot bullying in the posts themselves.

This is easier said than done.

'WE ARE IN A PIVOTAL MOMENT. WE WANT TO LEAD THE INDUSTRY IN THIS FIGHT.'

—ADAM MOSSERI

When engineers want to teach a machine to perform a task, they start by building a training set—in lay terms, a collection of material to help the machine understand the rules of its new job. In this case, it starts with human moderators sorting through hundreds of thousands of pieces of content and deciding whether they contain bullying. They label them and feed the examples into what is known as a classifier, which absorbs them like a preternatural police dog that is then set loose to sniff out that material. Of course, these initial examples can't cover everything a classifier encounters in the wild. But as it flags content, and as human moderators assess whether that was the correct call, it learns from the additional examples. Ideally, with the help of engineers tweaking its study habits, it gets better over time.

Today, there are three bullying classifiers scanning content on Instagram: one trained to analyze text, one photos and one videos. They're live, and they're flagging content by the hour. Yet they are also in "pretty early days," as lead engineer Yoav Shapira puts it. In other words, they're missing a lot of bullying. The mission of his team is to change that.

One reason this is so challenging—compared with training a machine to seek out content like nudity—is that it's much easier to recognize when someone in a photo is not wearing pants than it is to recognize the broad array of behavior that might be considered bullying. Studies of cyberbullying vary wildly in their conclusions on how many people have experienced it, from as low as 5% to as high as 72%, in part because no one agrees on precisely what it is. "What makes bullying so hard to tackle is that the definition is so different to individuals," says Newton. And engineers need a clear sense of what qualifies and what doesn't in order to build a solid training set.

Bullying on Instagram has changed over time. There is plenty of what one might call old-fashioned bullying: according to Instagram's own research, mean comments, insults and threats are most common. Some of this is easy to catch. Instagram's text classifier, for example, has been well trained to look for strings of words like "You ugly ass gapped tooth ass bitch" and "Your daughter is a slag." But slang changes over time and across cultures. And catching aggressive behavior requires comprehending full sentences, not just keywords. Consider the difference between "I'm coming over later" and "I'm coming over later no matter what you say."

Users also find themselves victimized by bullies who go beyond words. On Instagram, there are so-called hate pages, anonymous accounts dedicated to impersonating or mocking people. A boyfriend might tag an ex in posts that show him with other girls; a girl might tag some friends in a post and pointedly exclude someone. Others will take a screenshot of someone's photo, alter it and reshare it, or just mock it in a group chat. There's repeated contact that mimics stalking, like putting the same emoji on each photo someone posts. Many teens have embarrassing photos or videos of themselves shared without their consent, or find themselves the subject of "hot or not" votes (much like the rankings Mark Zuckerberg set up as a Harvard undergraduate on a pre-Facebook website called Facemash). "There's nothing in bullying," Shapira says, "that is super easy."

As it works to develop effective AI, Instagram has been surveying thousands of users in the hope of better understanding all the forms that bullying can take. The company's broad working definition of bullying is content intended to "harass or shame" others, but Shapira's team has broken this down into seven subcategories: insults, shaming, threats, identity attacks, disrespect, unwanted contact and betrayals. The grand plan is to build artificial intelligence that can understand each concept. "It's much more costly from an engineering perspective," Shapira says, "but it's much better at solving the problem."

Because bullying can be contextual—hinging on an inside joke or how well two people know each other—Instagram's engineers are also researching ways to use account behavior to separate the bad from the benign. The word *ho,* for example, might be classified as bullying when a man says it to a woman but not when a woman uses it to refer to a friend. Similarly, if someone says "Awesome picture" once, that might be a compliment. If they say it on every photo a person posts, that looks suspicious. Engineers are capitalizing on signals that help reveal relationships: Do two accounts tag each other a lot? Has one ever blocked the other? Does there seem to be a coordinated pile-on, like "Go to [someone's handle] and spam this picture on their dms"?

When it comes to photos and videos, the classifiers have had less practice and are less advanced. But some guideposts have emerged. For example, a split screen is often a sign of bullying, especially if a machine detects that one side shows a human and the other an animal. So is a photo of three people with a big red X across someone's face. Filters can help signal a benign post: people don't tend to pretty up their victimizing. The team is also learning to understand factors like posture. It's likely a photo was taken without consent if it appears to be an "upskirt" shot. If one person is standing and another is in the posture of a victim, that's a red flag.

Every week, researchers write up a report of their findings, and almost every week there's a new form of bullying that engineers hadn't thought to look for, Shapira says. But with the help of teams and resources from Facebook, employees who work on well-being believe they can not only master challenges that have long eluded experts—like building algorithms that understand sarcasm—but also figure out how to use AI to find newfangled phenomena like "intentional FOMO" and even accounts that torment middle schoolers about their necks.

There are a lot of numbers Instagram won't share, including how many of its roughly 1,200 employees, or Facebook's 37,700, are working on bullying. It also won't share the error rate of the classifiers that are currently live or the amount of content they're flagging for moderators. These days, the aspirations surrounding AI are often unmatched by the reality. But faith in the technology is high. "It's going to get much better," Shapira says, "over the next year or two."

'TECHNOLOGY ISN'T INHERENTLY GOOD OR BAD IN THE FIRST PLACE. IT JUST IS.'

—ADAM MOSSERI

MOSSERI INHERITED one of the most powerful perches in social media at a tough time for the industry. We sit down to talk about bullying in mid-May, in an airy room at Instagram's San Francisco office. It's his first on-the-record interview in the U.S., and it's coming shortly after the White House launched a tool inviting people who feel they've been censored by social-media companies to share their stories with the President. A few days before that, one of Facebook's co-founders had called for the breakup of the company.

Minefields are everywhere, but Mosseri is used to tricky terrain. After joining Facebook as a designer in 2008, he went on to help establish the team that oversees News Feed, the scroll of posts on everyone's home page that has been an epicenter of controversy for the company. When Facebook was manipulated by trolls and foreign meddlers during the 2016 election, News Feed is where it happened. In the wake of that, Mosseri established what he named the Integrity team and spent his time overseeing development of AI tools meant to root out complex ills like misinformation. In the process, he became close to Zuckerberg, and in early 2018 he was tapped to become Instagram's head of product, a post he soon leveraged into the top job.

As Instagram improves its definition of bullying, Mosseri believes the company will set new standards for using AI to squash it, developing practices that other platforms may even adopt. In the meantime, he's focused on finding ways to get Instagram's vast army of users to help combat this problem themselves, with the assistance of machines. "People often frame technology and people in opposition," he says. "My take is that people and technology can and should work in tandem."

Two new features Instagram is soon rolling out embody this approach. One is a comment warning. When someone decides to comment on a post, if Instagram's bullying classifier detects even "borderline" content, it will give that user a prompt, encouraging them to rethink their words. "The idea is to give you a little nudge and say, 'Hey, this might be offensive,' without blocking you from posting," says Francesco Fogu, a designer who works on well-being. The second feature is Restrict, a clandestine version of blocking. While users can tell when they've been blocked, they won't know if they've been restricted. Bullying victims will have the power to review comments from such accounts before anyone else sees them. They can approve, delete or forever leave them in a pending state: invisible to all but the bully. They'll have similar power over direct messages. And if the bully tries to tag that user in a public post, Instagram won't help by auto-completing the handle. The bully will have to know the username and type it exactly. All this, Fogu says, will "make it harder for bullies to bully others."

Instagram has also added a step to product development. Before anything is launched, it is now vetted for all the ways it could be "weaponized," including for bullying. If a feature is being abused frequently, Newton, the head of public policy, says the company will consider taking it away, even if it's popular.

This all might add up to Instagram doing more than any other company to fight this battle, but there are potential solutions it hasn't embraced. In interviews for this article, no issue came up more than user anonymity. It's what allows teenagers to run so-called confession or roast accounts dedicated to spreading gossip and skewering peers. It makes it easier for people to impersonate others. It encourages bad behavior. When asked if Instagram would consider requiring people to use their real names, as Facebook does, Mosseri doesn't rule it out, but he pushes back on the idea—one that would surely cut into Instagram's user base.

"People do tend to behave better when there's accountability, and part of accountability is that people know who you are, but a lot of good also comes from anonymity," he says. "People's identities are complicated." He, for one, has two children who are far too young to use Instagram (though people disobey the rule, kids under 13 aren't allowed), so he runs accounts for them that family members follow. And young people who are exploring their gender or sexual identities can do so without feeling exposed, he says. "The trick here is how do we make sure we address the incentives, address the problems," Mosseri says, "but also don't undo all the good."

THE PRESSURE to take bullying seriously has increased as scrutiny of tech companies has amped up, but also as research has shown it's not just some inconsequential rite of passage. For many teenagers, Instagram will be their first contact with social media, and they're arriving more lonely and depressed than previous generations were at their age. Sure, sometimes bullying is just annoying. But it has also been linked to self-harm, suicide and suicidal thoughts, as well as anxiety that can trail people into adulthood.

While things like name-calling are representative of bullying on the platform, there are isolated cases that are far more serious. In 2015, 12-year-old Kennis Cady died after falling into a coma while trying to hang herself in her East Rochester, N.Y., bedroom. In the investigation that followed, several students said they had seen or heard about an account on Instagram that two classmates had set up, one that allegedly pretended to be Kennis' and posted mocking "facts" about her. Investigators suspected Kennis was trying to delete the account before she died, though they never found it. Michaela Cady, her mother, says she "wholeheartedly believes" that bullying on social media contributed to her daughter's mental state. "She was withdrawn," Cady says. "She just seemed sad."

3

There are three classifiers currently seeking bullying on Instagram, one scanning text, one photos and one videos

7

Instagram is working to build AI to recognize seven subcategories of bullying, including shaming, disrespect and betrayal

13

Though some young people flout the rule, Instagram limits users to ages 13 and older

Before Mosseri sat down with TIME, the only other on-the-record interviews he had done came amid uproar earlier this year over a teenage girl in the United Kingdom who, according to her father, viewed material on topics like suicide and depression on Instagram before killing herself. Not by coincidence, the platform banned graphic images of self-harm in February.

Echoing other social-media power brokers, like Twitter CEO Jack Dorsey, Mosseri says the company is reassessing the "core incentives" that lead to angst on the platform. People do awful things—like tearing others down—in a quest for likes, so Instagram is experimenting with hiding that count entirely.

Calls to break up Big Tech have abounded on cable news and the campaign trail, including demands that Facebook spin off Instagram. Mosseri's response is that "the biggest downside" would be losing access his team has to AI and expertise they're able to use to address issues like bullying. He believes Instagram's massive scale is an asset too. Yes, if the platform were 10,000 times smaller and moderators reviewed each post and message, it might be easier to eradicate bullying, but that would present privacy problems. And at this size, the company is primed to better understand it and find technical means of countering it, he argues.

What it comes down to, Mosseri says, is whether one thinks connecting people—and even the Internet itself—is a net benefit for humanity. "Technology isn't inherently good or bad in the first place. It just is," he says. "And social media, as a type of technology, is often an amplifier. It's on us to make sure we're amplifying the good and not amplifying the bad."

Instagram didn't invent bullying. It's a problem essentially anywhere that people congregate online. Adults, from the President on down, are normalizing online abuse. Experts say that as a society, we are failing to teach kids how the Internet works before setting them loose on it. And they say that the fight against bullying can't be waged by tech companies alone: there needs to be buy-in from parents, schools and kids themselves.

Speaking from Texas, where her family moved after Kennis died, Michaela Cady says she does not blame Instagram for what happened. Kennis, like most kids who are bullied online, was also bullied in real life. But she does think that social media is too prevalent in society, that the solution isn't better technology but less of it. At this point, social media is so central to life that it's not realistic to tell kids who are bullied to just turn off their phones or delete their accounts. Yet, Cady says, there needs to be more space between the online and offline worlds. "With social media," she says, "there's just no escape." □