How Photos of Your Kids Are Powering Surveillance Technology

Millions of Flickr images were sucked into a database called MegaFace. Now some of those faces may have the ability to sue.

By Kashmir Hill and Aaron Krolik

One day in 2005, a mother in Evanston, Ill., joined Flickr. She uploaded some pictures of her children, Chloe and Jasper. Then she more or less forgot her account existed.

Years later, their faces are in a database that's used to test and train some of the most sophisticated artificial intelligence systems in the world.

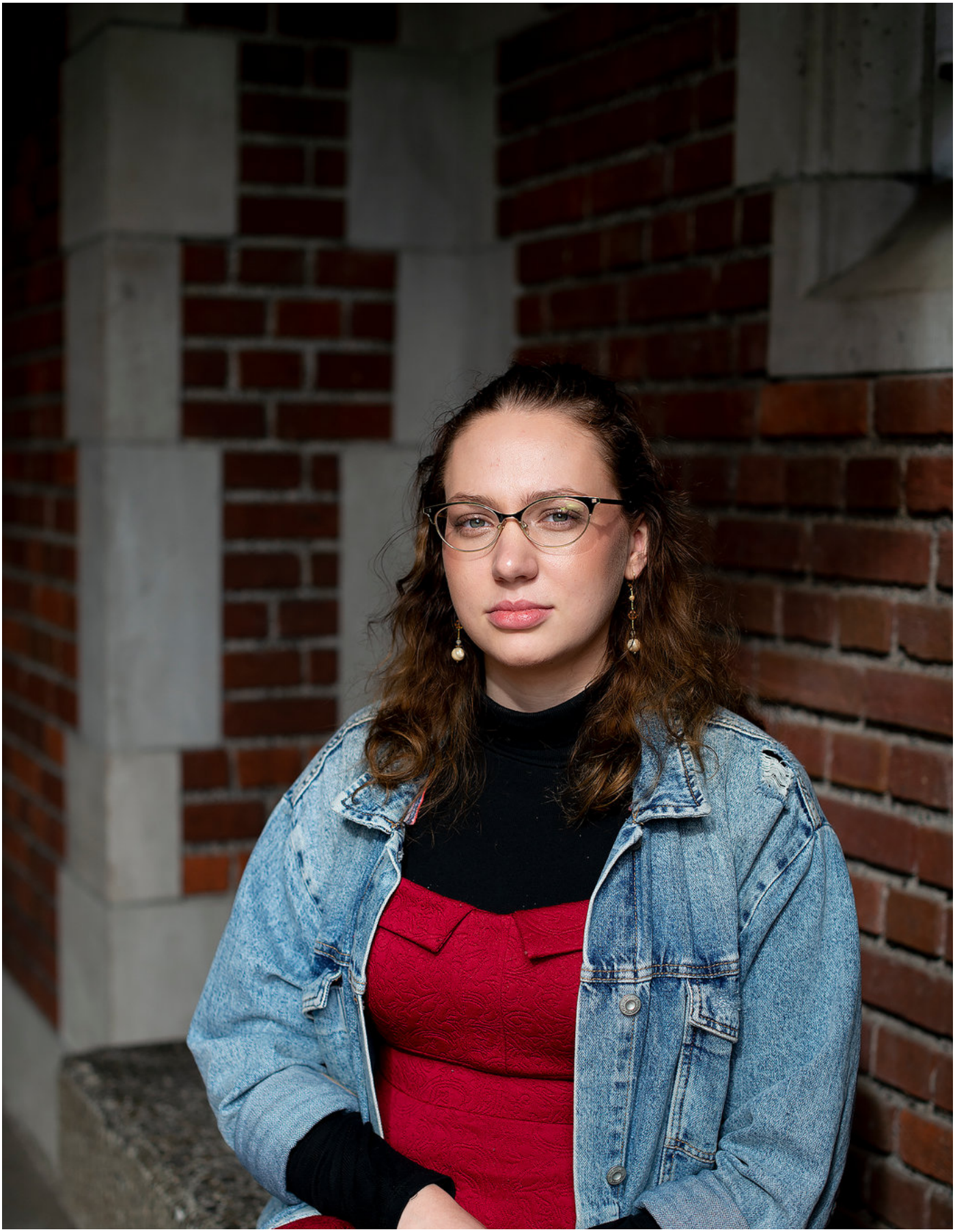
A selection of images from the MegaFace database.

The pictures of Chloe and Jasper Papa as kids are typically goofy fare: grinning with their parents; sticking their tongues out; costumed for Halloween. Their mother, Dominique Allman Papa, uploaded them to Flickr after joining the photo-sharing site in 2005.

None of them could have foreseen that 14 years later, those images would reside in an unprecedentedly huge facial-recognition database called MegaFace. Containing the likenesses of nearly 700,000 individuals, it has been downloaded by dozens of companies to train a new generation of face-identification algorithms, used to track protesters, surveil terrorists, spot problem gamblers and spy on the public at large.

"It's gross and uncomfortable," said Mx. Papa, who is now 19 and attending college in Oregon. "I wish they would have asked me first if I wanted to be part of it. I think artificial intelligence is cool and I want it to be smarter, but

generally you ask people to participate in research. I learned that in high school biology."



Chloe Papa Amanda Lucier for The New York Times

By law, most Americans in the database don't need to be asked for their permission — but the Papas should have been.

As residents of Illinois, they are protected by one of the strictest state privacy laws on the books: the Biometric Information Privacy Act, a 2008 measure that imposes financial penalties for using an Illinoisan's fingerprints or face scans without consent.

Those who used the database — companies including Google, Amazon, Mitsubishi Electric, Tencent and SenseTime — appear to have been unaware of the law, and as a result may have huge financial liability, according to several lawyers and law professors familiar with the legislation.

How MegaFace was born

How did the Papas and hundreds of thousands of other people end up in the database? It's a roundabout story.

In the infancy of facial-recognition technology, researchers developed their algorithms with subjects' clear consent: In the 1990s, universities had volunteers come to studios to be photographed from many angles. Later, researchers turned to more aggressive and surreptitious methods to gather faces at a grander scale, tapping into surveillance cameras in coffee shops, college campuses and public spaces, and scraping photos posted online.

According to Adam Harvey, an artist who tracks the data sets, there are probably more than 200 in existence, containing tens of millions of photos of

approximately one million people. (Some of the sets are derived from others, so the figures include some duplicates.) But these caches had flaws. Surveillance images are often low quality, for example, and gathering pictures from the internet tends to yield too many celebrities.

In June 2014, seeking to advance the cause of computer vision, Yahoo unveiled what it called "the largest public multimedia collection that has ever been released," featuring 100 million photos and videos. Yahoo got the images — all of which had Creative Commons or commercial use licenses — from Flickr, a subsidiary.

The database creators said their motivation was to even the playing field in machine learning. Researchers need enormous amounts of data to train their algorithms, and workers at just a few information-rich companies — like Facebook and Google — had a big advantage over everyone else.

"We wanted to empower the research community by giving them a robust database," said David Ayman Shamma, who was a director of research at Yahoo until 2016 and helped create the Flickr project. Users weren't notified that their photos and videos were included, but Mr. Shamma and his team built in what they thought was a safeguard.

They didn't distribute users' photos directly, but rather links to the photos; that way, if a user deleted the images or made them private, they would no longer be accessible through the database.

But this safeguard was flawed. The New York Times found a security vulnerability that allows a Flickr user's photos to be accessed even after they've been made private. (Scott Kinzie, a spokesman for SmugMug, which acquired Flickr from Yahoo in 2018, said the flaw "potentially impacts a very small number of our members today, and we are actively working to deploy an update as quickly as possible." Ben MacAskill, the company's chief operating officer, added that the Yahoo collection was created "years before our engagement with Flickr.")

Additionally, some researchers who accessed the database simply downloaded versions of the images and then redistributed them, including a team from the University of Washington. In 2015, two of the school's computer science professors — Ira Kemelmacher-Shlizerman and Steve Seitz — and their graduate students used the Flickr data to create MegaFace.

Containing more than four million photos of some 672,000 people, it held deep promise for testing and perfecting face-recognition algorithms.

Monitoring Uighurs and outing porn actors

Importantly to the University of Washington researchers, MegaFace included children like Chloe and Jasper Papa. Face-recognition systems tend to perform poorly on young people, but Flickr offered a chance to improve that with a bonanza of children's faces, for the simple reason that people love posting photos of their kids online.

In 2015 and 2016, the University of Washington ran the "MegaFace Challenge," inviting groups working on face-recognition technology to use the data set to test how well their algorithms were working.

The school asked people downloading the data to agree to use it only for "noncommercial research and educational purposes." More than 100 organizations participated, including Google, Tencent, SenseTime and NtechLab. In all, according to a 2016 university news release, "more than 300 research groups" have worked with the database. It has been publicly cited by researchers from Amazon and, according to Mr. Harvey, Mitsubishi Electric and Philips.

Some of these companies have been criticized for the way clients have deployed their algorithms: SenseTime's technology has been used to monitor the Uighur population in China, while NtechLab's has been used to out pornography actors and identify strangers on the subway in Russia.

SenseTime's chief marketing officer, June Jin, said that company researchers used the MegaFace database only for academic purposes. "Researchers have to use the same data set to ensure their results are comparable like-for-like," Ms. Jin wrote in an email. "As MegaFace is the most widely recognized database of its kind, it has become the de facto facial-recognition training and test set for the global academic and research community."

NtechLab spokesman Nikolay Grunin said the company deleted MegaFace after taking part in the challenge, and added that "the main build of our algorithm has never been trained on these images." Google declined to comment.

A spokeswoman for the University of Washington declined to make MegaFace's lead researchers available for interviews, saying they "have moved on to other projects and don't have the time to comment on this." Efforts to contact them individually were unsuccessful.

MegaFace's creation was financed in part by Samsung, Google's Faculty Research Award, and by the National Science Foundation/Intel.

In recent years, Ms. Kemelmacher-Shlizerman has sold a face-swapping image company to Facebook and advanced deep-fake technology by converting audio clips of Barack Obama into a realistic, synthetic video of him giving a speech. She is now working on a "moonshot project" at Google.

'What the hell? That is bonkers'

MegaFace remains publicly available for download. When The New York Times recently requested access, it was granted within a minute.

MegaFace doesn't contain people's names, but its data is not anonymized. A spokesman for the University of Washington said researchers wanted to honor the images' Creative Commons licenses. As a result, each photo

includes a numerical identifier that links back to the original Flickr photographer's account. In this way, The Times was able to trace many photos in the database to the people who took them.

"What the hell? That is bonkers," said Nick Alt, an entrepreneur in Los Angeles, when told his pictures were in the database, including photos he took of children at a public event in Playa Vista, Calif., a decade ago.

Mr. Alt's photos, with a selection of images from MegaFace.

"The reason I went to Flickr originally was that you could set the license to be noncommercial. Absolutely would I not have let my photos be used for machine-learning projects. I feel like such a schmuck for posting that picture. But I did it 13 years ago, before privacy was a thing."

Another subject, who asked to be identified as J., is now a 15-year-old high school sophomore in Las Vegas. Photos of him as a toddler are in the MegaFace database, thanks to his uncle's posting them to a Flickr album after a family reunion a decade ago. J. was incredulous that it wasn't illegal to put him in the database without his permission, and he is worried about the repercussions.

Since middle school, he has been part of an Air Force Association program called CyberPatriot, which tries to steer young people with programming skills toward careers in cybersecurity. "I'm very protective of my digital footprint because of it," he said. "I try not to post photos of myself online. What if I decide to work for the N.S.A.?"

For J., Mr. Alt and most other Americans in the photos, there is little recourse. Privacy law is generally so permissive in the United States that companies are free to use millions of people's faces without their knowledge to power the spread of face-recognition technology. But there is an exception.

In 2008, Illinois passed a prescient law protecting the "biometric identifiers and biometric information" of its residents. Two other states, Texas and Washington, went on to pass their own biometric privacy laws, but they aren't as robust as the one in Illinois, which strictly forbids private entities to collect, capture, purchase or otherwise obtain a person's biometrics — including a scan of their "face geometry" — without that person's consent.

"Photos themselves are not covered by the Biometric Information Privacy Act, but the scan of the photos should be. The mere use of biometric data is a violation of the statute," said Faye Jones, a law professor at the University of Illinois. "Using that in an algorithmic contest when you haven't notified people is a violation of the law."

Illinois residents like the Papas whose faceprints are used without their permission have the right to sue, said Ms. Jones, and are entitled to $1,000 per use, or $5,000 if the use was "reckless." The Times attempted to measure how many people from Illinois are in the MegaFace database; one approach, using self-reported location information, suggested 6,000 individuals, and another, using geotagging metadata, indicated as many as 13,000.

Their biometrics have likely been processed by dozens of companies. According to multiple legal experts in Illinois, the combined liability could add up to more than a billion dollars, and could form the basis of a class action.

"We have plenty of ambitious class-action lawyers here in Illinois," said Jeffrey Widman, the managing partner at Fox Rothschild in Chicago. "The law's been on the books in Illinois since 2008 but was basically ignored for a decade. I guarantee you that in 2014 or 2015, this potential liability wasn't on anyone's radar. But the technology has now caught up with the law."

A $35 billion case against Facebook

It's remarkable that the Illinois law even exists. According to Matthew Kugler,

a law professor at Northwestern University who has researched the Illinois act, it was inspired by the 2007 bankruptcy of a company called Pay by Touch, which had the fingerprints of many Americans, including Illinoisans, on file; there were worries that it could sell them during its liquidation.

No one from the technology industry weighed in on the bill, according to legislative and lobbying records.

"When the law was passed, no one who is now concerned about it was thinking about the issue," Mr. Kugler said. Silicon Valley is aware of the law now. Bloomberg News reported in April 2018 that lobbyists for Google and Facebook were trying to weaken its provisions.

More than 200 class-action lawsuits alleging misuse of residents' biometrics have been filed in Illinois since 2015, including a $35 billion case against Facebook for using face recognition to tag people in photos. That lawsuit gained momentum in August, when the United States Court of Appeals for the Ninth Circuit rejected the company's arguments that the people did not suffer "concrete harm."

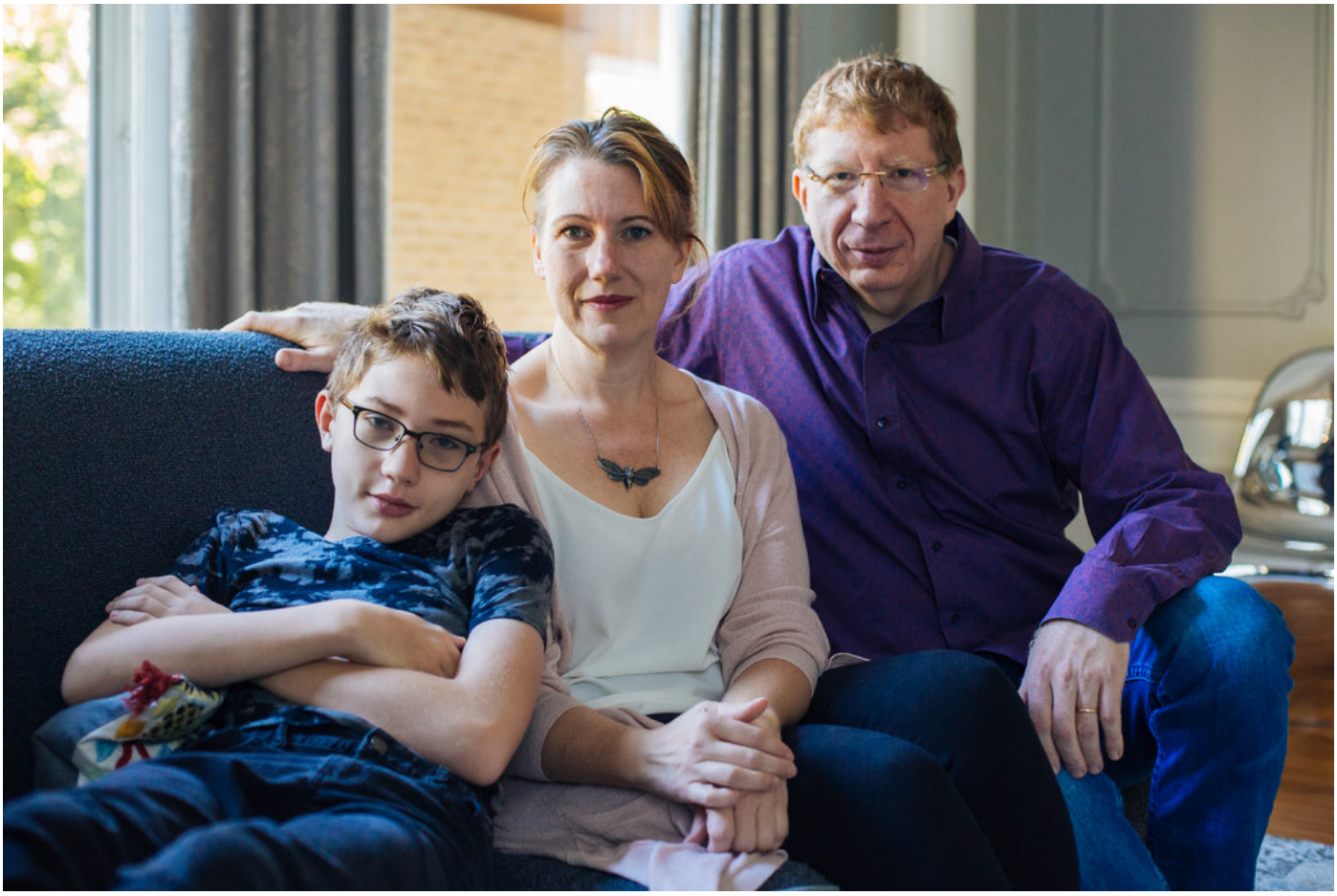
In recent years, technology companies have been treading more lightly in states with biometric legislation. When Google released a feature in 2018 that matched selfies to famous works of art, people in Illinois and Texas couldn't use it. And Google's Nest security cameras don't offer an otherwise standard feature for recognizing familiar faces in Illinois.

"It's creepy that you found me. I always lived with the philosophy that what I put out there was public, but I couldn't have imagined this," said Wendy Piersall, a publisher and City Council member in Woodstock, Ill., whose photos, along with those of her three children, were in the MegaFace database.

"We can't use the fun art app; why are you using our kids' faces to test your software?" she added. "My photos there are geotagged to Illinois. It's not

hard to figure out where these pictures were taken. I'm not a sue-happy person, but I would cheer someone else on to go after this."

Privacy nihilism



Dominique and George Papa with their son Jasper at their home in Evanston, Ill., earlier this month. Taylor Glascock for The New York Times

Some of the Illinois lawsuits have been settled or dismissed, but most are active, and Mr. Kugler, the Northwestern law professor, noted that basic legal questions remained unanswered. It's unclear what the legal liability would be for a company that takes photos uploaded in Illinois but processes the facial data in another state, or even another country.

"Defendants are going to be creative in searching for arguments, because no one wants to be stuck holding this expensive hot potato," he said.

A spokesman for Amazon Web Services said its use of the data set was

“compliant with B.I.P.A.,” while declining to explain how. Mario Fante, a spokesman for Philips, wrote in an email that the company “was never aware of any Illinois residents included in the above-mentioned data set.”

Victor Balta, a spokesman for the University of Washington, said, “All uses of photos in the researchers’ database are lawful. The U.W. is a public research university, not a private entity, and the Illinois law targets private entities.”

Some of the Illinoisans we found in MegaFace and contacted were indifferent about the use of their faces.

“I do know that when you upload information online, it can be used in unexpected ways, so I guess I’m not surprised,” said Chris Scheufele, a web developer in Springfield. “When you upload information to the internet and make it available for public consumption, you should expect it to be scraped.”

What about the subjects of his photos? Mr. Scheufele laughed. “I haven’t talked to my wife about it,” he said.

“Privacy nihilism” is an increasingly familiar term for giving up on trying to control data about oneself in the digital era. What happened to Chloe Papa could, depending on your perspective, argue for extreme vigilance or total resignation: Who could have possibly predicted that a snapshot of a toddler in 2005 would contribute, a decade and a half later, to the development of bleeding-edge surveillance technology?

“We have become accustomed to trading convenience for privacy, so that has dulled our senses as to what is happening with all the data gathered about us,” said Ms. Jones, the law professor. “But people are starting to wake up.”