KELVIN CHAN/AP/SHUTTERSTOCK

The Metropolitan Police in London used facial-recognition cameras to scan for wanted people in February.

BEATING BIOMETRIC BIAS

Facial recognition is improving – but the bigger issue is how it's used. **By Davide Castelvecchi**

When London's Metropolitan Police tested real-time facial-recognition technology between 2016 and 2019, they invited Daragh Murray to monitor some of the trials from a control room inside a police van.

"It's like you see in the movies," says Murray, a legal scholar at the University of Essex in Colchester, UK. As cameras scanned passers-by in shopping centres or public squares, they fed images to a computer inside the van. Murray and police officers saw the software draw rectangles around faces as it identified them in the live feed. It then extracted key features and compared them to those of suspects from a watch list. "If there is a match, it pulls an image from the live feed, together with the image from the watch list." Officers then reviewed the match and decided whether to rush out to stop the 'suspect' and, occasionally, arrest them.

Scotland Yard, as the headquarters of the London police force is sometimes known, had commissioned Murray and his University of Essex colleague Pete Fussey, a sociologist, to conduct an independent study of its dragnet. But their results[1], published in July 2019, might not have been quite what the law-enforcement agency had hoped for.

Fussey and Murray listed a number of ethical and privacy concerns with the dragnet, and questioned whether it was legal at all. And they queried the accuracy of the system, which is sold by Tokyo-based technology giant NEC. The software flagged 42 people over 6 trials that the researchers analysed; officers dismissed 16 matches as 'non-credible' but rushed out to stop the others. They lost 4 people in the crowd, but still stopped 22: only 8 turned out to be correct matches.

The police saw the issue differently. They said the system's number of false positives was tiny, considering the many thousands of faces that had been scanned. (They didn't reply to *Nature*'s requests for comment for this article.)

The accuracy of facial recognition has improved drastically since 'deep learning' techniques were introduced into the field about a decade ago. But whether that means it's good enough to be used on lower-quality, 'in the wild' images is a hugely controversial issue. And questions remain about how to transparently evaluate facial-recognition systems.

In 2018, a seminal paper by computer scientists Timnit Gebru, then at Microsoft Research in New York City and now at Google in Mountain View, California, and Joy Buolamwini at the Massachusetts Institute of Technology in Cambridge found that leading facial-recognition software packages performed much worse at identifying the gender of women and people of colour than at classifying male, white faces[2].

Concerns over demographic bias have since been quoted frequently in calls for moratoriums or bans of facial-recognition software.

In June, the world's largest scientific computing society, the Association for Computing Machinery in New York City, urged a suspension of private and government use of facial-recognition technology, because of "clear bias based on ethnic, racial, gender, and other human characteristics", which it said injured the rights of individuals in specific demographic groups. Axon, a maker of body cameras worn by police officers across the United States, has said that facial recognition isn't accurate enough to be deployed in its products. Some US cities have banned the use of the technology in policing, and US lawmakers have proposed a federal moratorium.

Companies say they're working to fix the biases in their facial-recognition systems, and some are claiming success. But many researchers and activists are deeply sceptical. They argue that even if the technology surpasses some benchmark in accuracy, that won't assuage deeper concerns that facial-recognition tools are used in discriminatory ways.

More accurate but still biased

Facial-recognition systems are often proprietary and swathed in secrecy, but specialists say that most involve a multi-stage process (see 'How facial recognition works') using deep learning to train massive neural networks on large sets of data to recognize patterns. "Everybody who does face recognition now uses deep learning," says Anil Jain, a computer scientist at Michigan State University in East Lansing.

The first stage in a typical system locates one or more faces in an image. Faces in the feed from a surveillance camera might be viewed in a range of lighting conditions and from different angles, making them harder to recognize than in a standard passport photo, for instance. The algorithm will have been trained on millions of photos to locate 'landmarks' on a face, such as the eyes, nose and mouth, and it distils the information into a compact file, ranging from less than 100 bytes to a few kilobytes in size.

The next task is to 'normalize' the face, artificially rotating it into a frontal, well-illuminated view. This produces a set of facial 'features' that can be compared with those extracted from an existing database of faces. This will typically consist of pictures taken under controlled conditions, such as police mugshots. Because the feature representations are compact, structured files, a computer can quickly scan millions of them to find the closest match.

Matching faces to a large database — called one-to-many identification — is one of two main types of facial-recognition system. The other is one-to-one verification, the relatively simple task of making sure that a person matches their own photo. It can be applied to anything from unlocking a smartphone to passport control at national borders.

One measure of progress is the Face Recognition Vendor Test, an independent benchmarking assessment that the US National Institute of Standards and Technology (NIST) in Gaithersburg, Maryland, has been conducting for two decades. Dozens of laboratories, both commercial and academic, have voluntarily taken part in the latest round of testing, which began in 2018 and is ongoing. NIST measures the performance of each lab's software package on its own image data sets, which include frontal and profile police mugshots, and pictures scraped from the Internet. (The US technology giants Amazon, Apple, Google and Facebook have not taken part in the test.)

In reports released late last year, the NIST team described massive steps forward in the technology's performance during 2018, both for one-to-many searches[3] and for one-to-one verification[4] (see also go.nature.com/35pku9q). "We have seen a significant improvement in face-recognition accuracy," says Craig Watson, an electrical engineer who leads NIST's image group. "We know that's largely because of convolutional neural networks," he adds, a type of deep neural network that is especially efficient at recognizing images.

The best algorithms can now identify people from a profile image taken in the wild — matching it with a frontal view from the database — about as accurately as the best facial-recognition software from a decade ago could recognize frontal images, NIST found. Recognizing a face in profile "has been a long-sought milestone in face recognition research", the NIST researchers wrote.

"Systems are being brought to the wild without a proper evaluation of their performance."

But NIST also confirmed what Buolamwini and Gebru's gender-classification work suggested: most packages tended to be more accurate for white, male faces than for people of colour or for women[5]. In particular, faces classified in NIST's database as African American or Asian were 10–100 times more likely to be misidentified than those classified as white. False positives were also more likely for women than for men.

This inaccuracy probably reflects imbalances in the composition of each company's training database, Watson says — a scourge that data scientists often describe as 'garbage in, garbage out'. Still, discrepancies varied between packages, indicating that some companies might have begun to address the problem, he adds.

NEC, which supplies Scotland Yard's software, noted that in NIST's analysis, it was "among a small group of vendors where false positives based on demographic differentials were undetectable", but that match rates could be compromised by outdoor, poorly lit or grainy images.

False faces

One-to-one verification, such as recognizing the rightful owner of a passport or smartphone, has become extremely accurate; here, artificial intelligence is as skilful as the sharpest-eyed humans. In this field, cutting-edge research focuses on detecting malevolent attacks. The first facial-recognition systems for unlocking phones, for example, were easily fooled by showing the phone a photo of the owner, Jain says; 3D face recognition does better. "Now the biggest challenge is very-high-quality face masks." In one project, Jain and his collaborators are working on detecting such impersonators by looking for skin texture.

But one-to-many verification, as Murray found, isn't so simple. With a large enough watch list, the number of false positives flagged up can easily outweigh the true hits.

This is a problem when police must make quick decisions about stopping someone. But mistakes also occur in slower investigations. In January, Robert Williams was arrested at his house in Farmington Hills, Michigan, after a police facial-recognition system misidentified him as a watch thief on the basis of blurry surveillance footage of a Black man, which it matched to his driving licence. The American Civil Liberties Union (ACLU), a non-profit organization in New York City, filed a complaint about the incident to Detroit police in June, and produced a video in which Williams recounts what happened when a detective showed him the surveillance photos on paper. "I picked that paper up, held it next to my face and said, 'This is not me. I hope y'all don't think all Black people look alike.' And then he said: 'The computer says it's you,'" Williams said. He was released after being detained for 30 hours. ACLU attorney Phil Mayor says the technology should be banned. "It doesn't work, and even when it does work, it remains too dangerous a tool for governments to use to surveil their own citizens for no compelling return," he says.

Shortly after the ACLU complaint, Detroit police chief James Craig acknowledged that the software, if used by itself, would misidentify cases "96% of the time". Citing concerns over racial bias and discrimination, at least 11 US cities have banned facial recognition by public authorities in the past 18 months. But Detroit police still use the technology. In late 2019, the force adopted policies to ban live camera surveillance and to use the software only on still images and as part of criminal investigations; Williams was arrested before the policy went into practice, Craig said in June. (He did not respond to *Nature*'s requests for comment.)

Other aspects of facial analysis, such as trying to deduce someone's personality on

the basis of their facial expressions, are even more controversial. Researchers have shown this doesn't work[6] – even the best software can only be trained on images tagged with other people's guesses. But companies around the world are still buying unproven technology that assesses interview candidates' personalities on the basis of videos of them talking.

Nuria Oliver, a computer scientist based in Alicante, Spain, says that governments should regulate the use of facial recognition and other potentially useful technologies to prevent abuses (see page 350). "Systems are being brought to the wild without a proper evaluation of their performance, or a process of verification and reproducibility," says Oliver, who is co-founder and vice-president of a regional network called the European Laboratory for Learning and Intelligent Systems.

Persistent problems

Some proposals for regulation have called for authorities to establish accuracy standards and require that humans review any algorithm's conclusions. But a standard based on, say, passing NIST benchmarks is much too low a bar on its own to justify deploying the technology, says Deborah Raji, a technology fellow in Ottawa with the Internet foundation Mozilla who specializes in auditing facial-recognition systems.

This year, Raji, Buolamwini, Gebru and others published another paper on the performance of commercial systems, and noted that although some firms had improved at classifying gender across lighter- and darker-skinned faces, they were still worse at guessing a person's age from faces with darker skin[7]. "The assessment process is incredibly immature. Every time we understand a new dimension to evaluate, we find out that the industry is not performing at the level that it thought it did," Raji says. It is important, she says, that companies disclose more about how they test and train their facial-recognition systems, and consult with the communities in which the technology will be used.

Technical standards cannot stop facial-recognition systems from being used in discriminatory ways, says Amba Kak, a legal scholar at New York University's AI Now Institute. "Are these systems going to be another tool to propagate endemic discriminatory practices in policing?" Human operators often end up confirming a system's biases rather than correcting it, Kak adds. Studies such as the Scotland Yard external review show that humans tend to overestimate the technology's credibility, even when they see the computer's false match next to the real face. "Just putting in a clause 'make sure there is a human in the loop' is not enough," she says.

Kak and others support a moratorium on any use of facial recognition, not just because the technology isn't good enough yet, but also because there needs to be a broader discussion of how to prevent it from being misused. The technology will improve, Murray says, but doubts will remain over the legitimacy of operating a permanent dragnet on innocent people, and over the criteria by which people are put on a watch list.

Concerns about privacy, ethics and human rights will grow. The world's largest biometric programme, in India, involves using facial recognition to build a giant national ID card system called Aadhaar. Anyone who lives in India can go to an Aadhaar centre and have their picture taken. The system compares the photo with existing records on 1.3 billion people to make sure the applicant hasn't already registered under a different name. "It's a mind-boggling system," says Jain, who has been a consultant for it. "The beauty of it is, it ensures one person has only one ID." But critics say it turns non-card owners into second-class citizens, and some allege it was used to purge legitimate citizens from voter rolls ahead of elections.

And the most notorious use of biometric technology is the surveillance state set up by the Chinese government in the Xinjiang province, where facial-recognition algorithms are used to help single out and persecute people from religious minorities (see page 354).

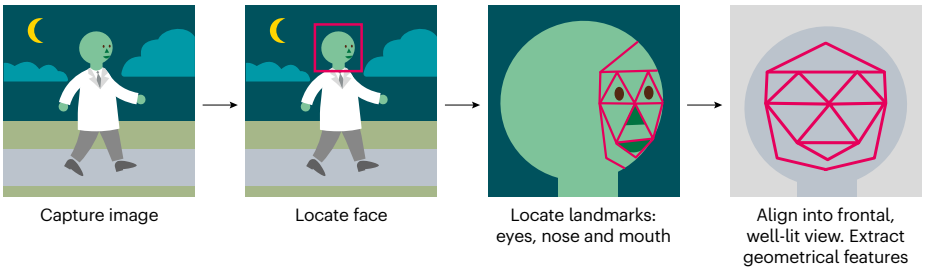
"At this point in history, we need to be a lot more sceptical of claims that you need ever-more-precise forms of public surveillance," says Kate Crawford, a computer scientist at New York University and co-director of the AI Now Institute. In August 2019, Crawford called for a moratorium on governments' use of facial-recognition algorithms (K. Crawford *Nature* **572**, 565; 2019).

Meanwhile, having declared its pilot project a success, Scotland Yard announced in January that it would begin to deploy live facial recognition across London.

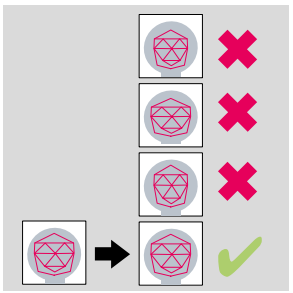
Davide Castelvecchi reports for *Nature* from London. Additional reporting by **Antoaneta Roussi and Richard Van Noorden**.

HOW FACIAL RECOGNITION WORKS

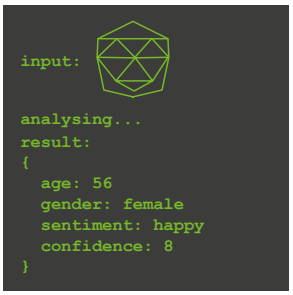
Facial-recognition systems analyse a face's geometry to create a faceprint — a biometric marker that can be used to recognize or identify a person. Another use is facial analysis, which tries to classify a face according to labels such as gender, age, ethnicity or emotion.



Examples:
- Unlock a smartphone
- Travel through a passport gate
- Verify school or work attendance



Examples:
- Scan crowd until 'hit' found against watch list
- Match person of interest against vast database



Examples:
- Assess person's age or gender
- Assess person's emotional state (tests are controversial and less reliable than facial recognition).

1. Fussey, P. & Murray, D. *Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology* (Univ. Essex, 2019).
2. Buolamwini, J. & Gebru, T. *Proc. Mach. Learn. Res.* **81**, 77–91 (2018).
3. Grother, P., Ngan, M. & Hanaoka, K. *Face Recognition Vendor Test (FRVT) Part 2: Identification* (NIST, 2019).
4. Grother, P., Ngan, M. & Hanaoka, K. *Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification. Updated 10 September 2020* (NIST, 2020).
5. Grother, P., Ngan, M. & Hanaoka, K. *Face Recognition Vendor Test Part 3: Demographic Effects* (NIST, 2019).
6. Feldman Barrett, L., Adolphs, R., Marsella, S., Martinez, A. M. & Pollak, S. D. *Psychol. Sci. Public Interest* **20**, 1–68 (2019).
7. Raji, I. D. *et al* Preprint at https://arxiv.org/abs/2001.00964 (2020).