the past few years provide an imperative for developing and validating oximeters without a fundamental dependence on pigmentation. These studies also highlight the importance of carefully reconsidering the enrolment criteria suggested for calibration studies, so that the skin pigmentation of test participants is evenly balanced, and determined using objective measures.

Chetan Patil and **Mohammed Shahriar Arefin** are in the College of Engineering, Temple University, Philadelphia, Pennsylvania 19122, USA.
e-mails: c.patil@temple.edu; shahriar.arefin@temple.edu

1. Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E. & Valley, T. S. *N. Engl. J. Med.* **383**, 2477–2478 (2020).
2. Fawzy, A. *et al. JAMA Intern. Med.* **182**, 730–738 (2022).
3. Gottlieb, E. R., Ziegler, J., Morley, K., Rush, B. & Celi, L. A. *JAMA Intern. Med.* **182**, 849–858 (2022).
4. Valbuena, V. S. M. *et al. BMJ* **378**, e069775 (2022).
5. Wong, A.-K. I. *et al. JAMA Netw. Open* **4**, e2131674 (2021).
6. Ries, A. L., Prewitt, L. M. & Johnson, J. J. *Chest* **96**, 287–290 (1989).
7. Feiner, J. R., Severinghaus, J. W. & Bickler, P. E. *Anesth. Analg.* **105**, S18–S23 (2007).
8. Okunlola, O. E. *et al. Respir. Care* **67**, 252–257 (2022).
9. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. *N. Engl. J. Med.* **383**, 874–882 (2020).
10. Severinghaus, J. W. *Anesth. Analg.* **105**, S1–S4 (2007).
11. Jacques, S. L. *Phys. Med. Biol.* **58**, R37 (2013).
12. Mannheimer, P. D. *Anesth. Analg.* **105**, S10–S17 (2007).
13. Wang, L., Jacques, S. L. & Zheng, L. *Comput. Methods Programs Biomed.* **47**, 131–146 (1995).
14. Chatterjee, S. & Kyriacou, P. A. *Sensors* **19**, 789 (2019).
15. Boonya-ananta, T. *et al. Sci. Rep.* **11**, 2570 (2021).
16. Arefin, M. S., Dumont, A. P. & Patil, C. A. *Proc. SPIE* **11951**, 1195103 (2022).
17. Fine, J. *et al. Biosensors* **11**, 126 (2021).

The authors declare no competing interests.

In Retrospect

Racism in science

The unseen Black faces of AI algorithms

Abeba Birhane

An audit of commercial facial-analysis tools found that dark-skinned faces are misclassified at a much higher rate than are faces from any other group. Four years on, the study is shaping research, regulation and commercial practices.

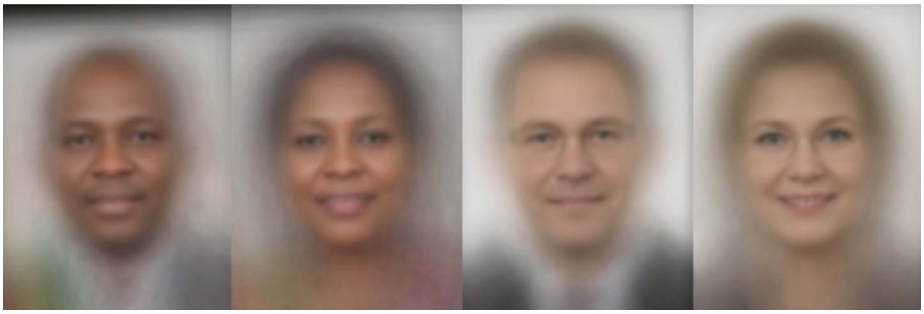
Data sets are essential for training and validating machine-learning algorithms. But these data are typically sourced from the Internet, so they encode all the stereotypes, inequalities and power asymmetries that exist in society. These biases are exacerbated by the algorithmic systems that use them, which means that the output of the systems is discriminatory by nature, and will remain problematic and potentially harmful until the data sets are audited and somehow corrected. Although this has long been the case, the first major steps towards overcoming the issue were taken only four years ago, when Joy Buolamwini and Timnit Gebru[1] published a report that kick-started sweeping changes in the ethics of artificial intelligence (AI).

As a graduate student in computer science, Buolamwini was frustrated that commercial facial-recognition systems failed to identify her face in photographs and video footage. She hypothesized that this was due, in part, to the fact that dark-skinned faces were not represented in the data sets that were used to train the computer programs she was studying. This insight led Buolamwini and her collaborator Gebru to undertake a systematic audit of commercial facial-analysis systems, and to demonstrate that such systems perform differently depending on the skin colour and gender of the person in the image. The work became known as the Gender Shades audit.

The authors began by using a skin-type classification system, approved by dermatologists, to assess the composition of two image banks, known as IJB-A and Adience, that were widely used at the time to train facial-recognition software. They found that individuals with light-coloured skin were the subject of 79.6% of the images in IJB-A and of 86.2% of those in Adience. This prompted Buolamwini and Gebru to compile their own set of images – one that offered a broader range of skin tones than did either of the existing options, as well as including similar numbers of men and women (commercial algorithms are typically not capable of dealing with non-binary classifications). To do so, they turned to photographs of politicians from countries with gender parity in their national parliaments. The resulting data set, known as the Pilot Parliaments Benchmark (Fig. 1), contains images of 1,270 individuals from Rwanda, Senegal, South Africa, Iceland, Finland and Sweden.

Buolamwini and Gebru then used their benchmark set to evaluate three commercial gender-classification systems developed by the technology companies Microsoft, Face++ and IBM. Rather than assessing the accuracy of these systems on the basis of gender or of skin type, the authors compared the performance of the classifiers on four intersectional groups that they termed darker female, darker male, lighter female and lighter male. They found that women with darker skin were the most likely to be misclassified, with a maximum classification error rate of 34.7%; by contrast, the maximum error rate for men with lighter skin was 0.8%. All three systems consistently showed poor accuracy for women with dark skin and performed substantially better on white men.

Impactful research isn't always understood and acknowledged at first glance, especially when it challenges conventional thinking. At the time of publication, Buolamwini and Gebru's paper was considered an outlier – not only in the field of computer vision (the study of how computers can be made to automate tasks



DR. JOY BUOLAMWINI (CC BY 4.0)

Figure 1 | A gender-balanced facial image bank with a range of skin tones. On realizing that dark-skinned faces were under-represented in the data sets of images that are used to train facial-recognition software, Buolamwini and Gebru[1] compiled their own data set using photographs of politicians from countries with gender parity in their national parliaments. This is a subset of 'average' faces made by blending many images from the full data set, which contains photographs of 1,270 individuals from Rwanda, Senegal, South Africa, Iceland, Finland and Sweden. Buolamwini and Gebru used their data set to show that three commercial gender-classification systems misclassified women with darker skin with an error rate that was much higher than that for men with lighter skin.

performed by the human visual system), but also in AI ethics. Since then, a lot has changed, and algorithmic auditing has rapidly become a crucial practice, prompting academic journals and conferences to highlight audit studies.

The downstream effect of the Gender Shades audit in research can also be found in curation practices for large-scale data sets. For instance, an initiative reported earlier this year suggests that faces in large image banks, such as the popular ImageNet (go.nature.com/3qukjkn), should be obscured to protect individuals' privacy[2]. The study showed that blurring or mosaicking faces in an image had little effect on the accuracy of software designed to recognize other elements of the image. But the authors also noted that this work must be done through crowdsourcing, rather than using commercial software, to avoid the racial bias revealed by the Gender Shades study.

Although there was resistance to Buolamwini and Gebru's paper at first, the vendors of the facial-recognition software that they audited eventually responded positively. IBM and Microsoft, for example, pledged to test their facial-recognition algorithms and diversify their training data sets (see, for example, go.nature.com/3rmbo17). Around a year after the paper was published, a follow-up audit found that Microsoft, IBM and Face++ had all succeeded in reducing the performance error of their facial-analysis products[3]. The most noteworthy improvement was a 30.4% reduction in the error with which the Face++ software recognized darker female faces in the Pilot Parliaments Benchmark set, with the Microsoft and IBM algorithms improving by 19.28% and 17.73%, respectively, on this task.

But none of these systems has yet overcome racial bias entirely, and many companies have discontinued or temporarily halted facial-recognition technologies. Evidence continues to emerge that AI models mistakenly associate images of Black people with animal classes such as 'gorilla' or 'chimpanzee' more often than they do for images of people who aren't Black[4].

The study also influenced how facial-analysis technology is regulated. In the United States, the 2019 Algorithmic Accountability Act authorized the Federal Trade Commission (the agency tasked with promoting and enforcing consumer protection) to regulate automated decision systems (go.nature.com/3xguff7). US cities such as San Francisco in California, Boston, Massachusetts, and Portland, Oregon, have banned the use of facial recognition by police, citing biased misidentification that disproportionately affects communities of colour. In Europe, civil-society organizations, activists and technologists have come together to call for a ban on facial-recognition analysis (go.nature.com/3qwzmnq) and on biometric technology in general (go.nature.com/3f7jrka). And the first draft of the European Union's Artificial Intelligence Act (go.nature.com/3dtgh4x), released in April 2021, indicates that real-time use of facial recognition in public places might be restricted.

Regulations and the risk of liability impel large corporations to change their practices, but even minimal regulations are being undermined (go.nature.com/3yb96kq). Although such deterrents can result in measurably improved outcomes, given the prevalence of facial-analysis technology, the changes that have the most long-term impact are likely to come from shifting public attitudes — something that I think Buolamwini and Gebru's study has influenced both directly and indirectly. The work even became the subject of the 2020 documentary film *Coded Bias* (go.nature.com/3fashnf). Unfortunately, the authors (like many other Black female scholars) have also been overlooked by mainstream media: a 2021 television segment on racial bias in facial-analysis technologies, for example, failed to recognize their work and that of their collaborators (go.nature.com/3satrp8).

> **"Buolamwini was frustrated that commercial facial-recognition systems failed to identify her face."**

Over the past few years, the conversation initiated by this work has shifted from a focus on the accuracy and performance of facial-recognition algorithms to larger and more-fundamental questions around surveillance technology. The question of accuracy becomes meaningless when this technology is used to supposedly measure internal behaviours from outward appearances. In fact, 'accurate' representation boils down to reducing these behaviours to outdated social stereotypes[5].

Algorithms that claim to detect emotions, predict gender or gauge someone's trustworthiness have been dubbed 'AI snake oil' by some (go.nature.com/3rh7cfp), because such sociocultural attributes cannot reliably be inferred from faces, expressions or gestures[6]. Others have called for a blanket ban on facial-recognition algorithms, saying that the technology resurrects the pseudosciences of physiognomy and phrenology[7].

The ImageNet data set, a large-scale collection of images that is considered the gold standard in computer vision, has had a pivotal role in positioning computer-vision research at the core of the 'deep-learning revolution' of the past decade. Facial-recognition technology has subsequently become mainstream and is prevalent in almost all social and public spaces, including concert venues, schools, airports, neighbourhoods and public squares — seriously undermining privacy and enabling worrying surveillance practices. Even if new algorithms are designed on the basis of diverse image sets such as the Pilot Parliaments Benchmark, they are still vulnerable to being used for inherently harmful and oppressive purposes, such as the surveillance of minority communities.

Facial-recognition technology has expanded into other fields of research, such as studies designed to predict facial characteristics from the analysis of DNA[8], and others that aim to automate medical diagnoses from images of faces alone[9]. Given the racial biases inherent in facial-recognition algorithms, these are concerning developments.

Amid what can feel like overwhelming public enthusiasm for new AI technologies, Buolamwini and Gebru instigated a body of critical work that has exposed the bias, discrimination and oppressive nature of facial-analysis algorithms. Their audit was ground-breaking four years ago, and remains an influential reference point to counter the rapid progress of this technology and the threat it poses.

Abeba Birhane is at the Mozilla Foundation, San Francisco, California 94105, USA, and in the School of Computer Science, University College Dublin, Dublin, Ireland.
e-mail: abeba@mozillafoundation.org

1. Buolamwini, J. & Gebru, T. *Proc. Mach. Learn. Res.* **81**, 77–91 (2018).
2. Yang, K., Yau, J. H., Fei-Fei, L., Deng, J. & Russakovsky, O. *Proc. Mach. Learn. Res.* **162**, 25313–25330 (2022).
3. Raji, I. D. & Buolamwini, J. In *Proc. 2019 AAAI/ACM Conference on AI, Ethics, and Society* 429–435 (Association for Computing Machinery, 2019).
4. Radford, A. *et al.* Preprint at https://arxiv.org/abs/2103.00020 (2021).
5. Birhane, A. *Artif. Life* **27**, 44–61 (2021).
6. Birhane, A. & Guest, O. *Women Gender Res.* **29**, 60–73 (2021).
7. Stark, L. & Hutson, J. *Fordham Intellect. Prop. Media Entertain. Law J.* https://doi.org/10.2139/ssrn.3927300 (2021).
8. Sero, D. *et al. Nature Commun.* **10**, 2557 (2019).
9. Thevenot, J., Bordallo López, M. & Hadid, A. *IEEE J. Biomed. Health Inform.* **22**, 1497–1511 (2018).

The author declares no competing interests.

Correction
The original version of this article contained a misspelling in the author's e-mail address, which has now been corrected.