for each sector. Producers of low-carbon goods should be allowed to document their actual emissions with audited data. The BCA should reflect the difference in carbon constraints between the imposing and targeted economies. For example, if aluminium in the imposing country faces an average compliance cost of $30 per tonne of carbon emissions, and imported aluminium is subject only to an average of $10 per tonne in its country of origin, then the $20 difference would be imposed as a BCA.

"Turn this incipient trade war into an opportunity to ratchet up climate ambition."

Ensure the process is fair. Equity, transparency and predictability are essential to legal durability. Countries should notify trade partners in advance and discuss the details with them. An independent arbiter could audit the plans before they are adopted and determine whether they are reasonable. The plan should include procedures for appealing.

The next few years will be a crucial time for both trade and climate policy. Trump plans to renegotiate the North American Free Trade Agreement (NAFTA), having already ended US participation in the Trans-Pacific Partnership (TPP). The United Kingdom and EU must rewrite their joint policies on trade and climate change. Nations will review their Paris pledges with a view to strengthening climate ambition. All these processes are parts of a larger puzzle.

As the pieces of the jigsaw fall into place, momentum must be sustained on climate change. Rather than prolonging the current spiral of tariff tit-for-tat, countries should rally and turn this incipient trade war into an opportunity to ratchet up climate ambition. ■

Michael A. Mehling *is deputy director of the Center for Energy and Environmental Policy Research, Massachusetts Institute of Technology, Cambridge, USA; and professor at the University of Strathclyde School of Law, Glasgow, UK.* **Harro van Asselt** *is senior research fellow at the Stockholm Environment Institute; and professor of climate law and policy, University of Eastern Finland Law School, Joensuu, Finland.* **Kasturi Das** *is professor of economics and international business, Institute of Management Technology, Ghaziabad, India.* **Susanne Droege** *is senior fellow, Global Issues Research Division, German Institute for International and Security Affairs, Berlin, Germany.*
e-mail: mmehling@mit.edu

Design AI so that it's fair

Identify sources of inequity, de-bias training data and develop algorithms that are robust to skews in data, urge **James Zou** and **Londa Schiebinger**.

When Google Translate converts news articles written in Spanish into English, phrases referring to women often become 'he said' or 'he wrote'. Software designed to warn people using Nikon cameras when the person they are photographing seems to be blinking tends to interpret Asians as always blinking. Word embedding, a popular algorithm used to process and analyse large amounts of natural-language data, characterizes European American names as pleasant and African American ones as unpleasant.

These are just a few of the many examples uncovered so far of artificial intelligence (AI) applications systematically discriminating against specific populations.

Biased decision-making is hardly unique to AI, but as many researchers have noted[1], the growing scope of AI makes it particularly important to address. Indeed, the ubiquitous nature of the problem means that we need systematic solutions. Here we map out several possible strategies.

SKEWED DATA

In both academia and industry, computer scientists tend to receive kudos (from publications to media coverage) for training ever more sophisticated algorithms. Relatively

Algorithms trained on biased data sets often recognize only the left-hand image as a bride.

LEFT: ISTOCK/GETTY; RIGHT: PRAKASH SINGH/AFP/GETTY

little attention is paid to how data are collected, processed and organized.

A major driver of bias in AI is the training data. Most machine-learning tasks are trained on large, annotated data sets. Deep neural networks for image classification, for instance, are often trained on ImageNet, a set of more than 14 million labelled images. In natural-language processing, standard algorithms are trained on corpora consisting of billions of words. Researchers typically construct such data sets by scraping websites, such as Google Images and Google News, using specific query terms, or by aggregating easy-to-access information from sources such as Wikipedia. These data sets are then annotated, often by graduate students or through crowdsourcing platforms such as Amazon Mechanical Turk.

Such methods can unintentionally produce data that encode gender, ethnic and cultural biases.

Frequently, some groups are over-represented and others are under-represented. More than 45% of ImageNet data, which fuels research in computer vision, comes from the United States[2], home to only 4% of the world's population. By contrast, China and India together contribute just 3% of ImageNet data, even though these countries represent 36% of the world's population. This lack of geodiversity partly explains why computer vision algorithms label a photograph of a traditional US bride dressed in white as 'bride', 'dress', 'woman', 'wedding', but a photograph of a North Indian bride as 'performance art' and 'costume'[2].

In medicine, machine-learning predictors can be particularly vulnerable to biased training sets, because medical data are especially costly to produce and label. Last year, researchers used deep learning to identify skin cancer from photographs. They trained their model on a data set of 129,450 images, 60% of which were scraped from Google Images[3]. But fewer than 5% of these images are of dark-skinned individuals, and the algorithm wasn't tested on dark-skinned people. Thus the performance of the classifier could vary substantially across different populations.

Another source of bias can be traced to the algorithms themselves.

A typical machine-learning program will try to maximize overall prediction accuracy for the training data. If a specific group of individuals appears more frequently than others in the training data, the program will optimize for those individuals because this boosts overall accuracy. Computer scientists evaluate algorithms on 'test' data sets, but usually these are random sub-samples of the original training set and so are likely to contain the same biases.

Flawed algorithms can amplify biases through feedback loops. Consider the case of statistically trained systems such as Google Translate defaulting to the masculine pronoun. This patterning is driven by the ratio of masculine pronouns to feminine pronouns in English corpora being 2:1. Worse, each time a translation program defaults to 'he said', it increases the relative frequency of the masculine pronoun on the web — potentially reversing hard-won advances towards equity[4]. The ratio of masculine to feminine pronouns has fallen from 4:1 in the 1960s, thanks to large-scale social transformations.

TIPPING THE BALANCE

Biases in the data often reflect deep and hidden imbalances in institutional infrastructures and social power relations. Wikipedia, for example, seems like a rich and diverse data source. But fewer than 18% of the site's biographical entries are on women. Articles about women link to articles about men more often than vice versa, which makes men more visible to search engines. They also include more mentions of romantic partners and family[5].

Thus, technical care and social awareness must be brought to the building of data sets for training. Specifically, steps should be taken to ensure that such data sets are diverse and do not under represent particular groups. This means going beyond convenient classifications —'woman/man', 'black/white', and so on — which fail to capture the complexities of gender and ethnic identities.

Some researchers are already starting to work on this (see *Nature* **558**, 357–360; 2018). For instance, computer scientists recently revealed that commercial facial recognition systems misclassify gender much more often when presented with darker-skinned women compared with lighter-skinned men, with an error rate of 35% versus 0.8% (ref. 6). To address this, the researchers curated a new image data set composed of 1,270 individuals, balanced in gender and ethnicity. Retraining and fine-tuning existing face-classification algorithms using these data should improve their accuracy.

To help identify sources of bias, we recommend that annotators systematically label the content of training data sets with standardized metadata. Several research groups are already designing 'datasheets'[7] that contain metadata and 'nutrition labels' for machine-learning data sets (http://data-nutrition.media.mit.edu/).

Every training data set should be accompanied by information on how the data were collected and annotated. If data contain information about people, ▶

then summary statistics on the geography, gender, ethnicity and other demographic information should be provided (see 'Image power'). If the data labelling is done through crowdsourcing, then basic information about the crowd participants should be included, alongside the exact request or instruction that they were given.

As much as possible, data curators should provide the precise definition of descriptors tied to the data. For instance, in the case of criminal-justice data, appreciating the type of 'crime' that a model has been trained on will clarify how that model should be applied and interpreted.

BUILT-IN FIXES

Many journals already require authors to provide similar types of information on experimental data as a prerequisite for publication. For instance, *Nature* asks authors to upload all microarray data to the open-access repository Gene Expression Omnibus — which in turn requires authors to submit metadata on the experimental protocol. We encourage the organizers of machine-learning conferences, such as the International Conference on Machine Learning, to request standardized metadata as an essential component of the submission and peer-review process. The hosts of data repositories, such as OpenML, and AI competition platforms, such as Kaggle, should do the same.

Lastly, computer scientists should strive to develop algorithms that are more robust to human biases in the data.

Various approaches are being pursued. One involves incorporating constraints and essentially nudging the machine-learning model to ensure that it achieves equitable performance across different subpopulations and between similar individuals[8]. A related approach involves changing the learning algorithm to reduce its dependence on sensitive attributes, such as ethnicity, gender, income — and any information that is correlated with those characteristics[9].

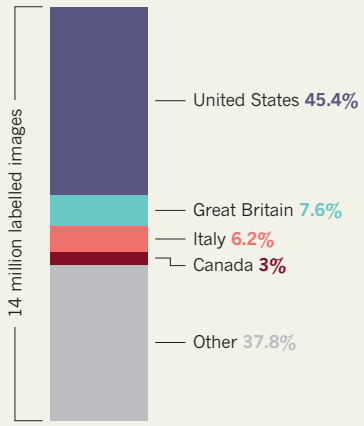
Such nascent de-biasing approaches are promising, but they need to be refined and evaluated in the real world.

An open challenge with these types of solutions, however, is that ethnicity, gender and other relevant information need to be accurately recorded. Unless the appropriate categories are captured, it's difficult to know what constraints to impose on the model, or what corrections to make. The approaches also require algorithm designers to decide a priori what types of biases they want to avoid.

A complementary approach is to use machine learning itself to identify and quantify bias in algorithms and data. We call this conducting an AI audit, in which the auditor is an algorithm that systematically probes the original machine-learning model to identify biases in both the model and the training data.

IMAGE POWER

Deep neural networks for image classification are often trained on ImageNet. The data set comprises more than 14 million labelled images, but most come from just a few nations.



SOURCE: REF 2

An example of this is our recent work using a popular machine-learning method called word embedding to quantify historical stereotypes in the United States. Word embedding maps each English word to a point in space (a geometric vector) such that the distance between vectors captures semantic similarities between corresponding words. It captures analogy relations, such as 'man' is to 'king' as 'woman' is to 'queen'. We developed an algorithm — the AI auditor — to query the word embedding for other gender analogies. This has revealed that 'man' is to 'doctor' as 'woman' is to 'nurse', and that 'man' is to 'computer programmer' as 'woman' is to 'homemaker'[1].

"Biases in the data often reflect deep and hidden imbalances in institutional infrastructures and social power relations."

Once the auditor reveals stereotypes in the word embedding and in the original text data, it is possible to reduce bias by modifying the locations of the word vectors. Moreover, by assessing how stereotypes have evolved, algorithms that are trained on historical texts can be debiased. Embeddings for each decade of US text data from Google Books from 1910 to 1990, reveal, for instance, shocking and shifting attitudes towards Asian Americans. This group goes from being described as 'monstrous' and 'barbaric' in 1910 to 'inhibited' and 'sensitive' in 1990 — with abrupt transitions after the Second World War and the immigration waves of the 1980s[10].

GETTING IT RIGHT

As computer scientists, ethicists, social scientists and others strive to improve the fairness of data and of AI, all of us need to think about appropriate notions of fairness. Should the data be representative of the world as it is, or of a world that many would aspire to? Likewise, should an AI tool used to assess potential candidates for a job evaluate talent, or the likelihood that the person will assimilate well into the work environment? Who should decide which notions of fairness to prioritize?

To address these questions and evaluate the broader impact of training data and algorithms, machine-learning researchers must engage with social scientists, and experts in the humanities, gender, medicine, the environment and law. Various efforts are under way to try to foster such collaboration, including the 'Human-Centered AI' initiative that we are involved in at Stanford University in California. And this engagement must begin at the undergraduate level. Students should examine the social context of AI at the same time as they learn about how algorithms work.

Devices, programs and processes shape our attitudes, behaviours and culture. AI is transforming economies and societies, changing the way we communicate and work and reshaping governance and politics. Our societies have long endured inequalities. AI must not unintentionally sustain or even worsen them. ■

James Zou *is assistant professor of biomedical data science and (by courtesy) of computer science and of electrical engineering, Stanford University, California, USA.* **Londa Schiebinger** *is the John L. Hinds Professor of History of Science and director of Gendered Innovations in Science, Health & Medicine, Engineering, and Environment, Stanford University, California, USA.*
e-mails: jamesz@stanford.edu; schieb@stanford.edu

1. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. *Adv. Neural Inf. Proc. Syst.* **2016**, 4349–4357 (2016).
2. Shankar, S. *et al.* Preprint at https://arxiv.org/abs/1711.08536 (2017).
3. Esteva, A. *et al. Nature* **542**, 115–118 (2017).
4. Schiebinger, L. *et al.* (eds) *Gendered Innovations in Science, Health & Medicine, Engineering and Environment, Engineering, Machine Translation* (2011–2015).
5. Wagner, C., Garcia, D., Jadidi, M. & Strohmaier, M. *Proc. 9th Int. AAAI Conf. Web Soc. Media* 454–463 (2015).
6. Buolamwini, J. & Gebru, T. *Proc. Mach. Learn. Res.* **81**, 1–15 (2018).
7. Gebru, T. *et al.* Preprint at https://arxiv.org/abs/1803.09010 (2018).
8. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. *Proc. 3rd Innov. Theor. Comp. Sci. Conf.* **2012**, 214–226 (2012).
9. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. *Proc. 30th Int. Conf. Int. Conf. Mach. Learn.* **28**, III-325–III-333 (2013).
10. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. *Proc. Natl Acad. Sci. USA* **115**, E3635–E3644 (2018).