AGENDA

Advancing economics in business

oxera

October 2020

The risks of using AI in business: artificial intelligence and real discrimination

Contact
Pascale Déchamps
Partner

Algorithms influence many aspects of our work and social lives. They affect what adverts we see, what shows we watch, and whether we get a job. As these tools become increasingly widespread, they pose new challenges to businesses. We look at concerns regarding the use of algorithms in areas where the role of computer programs and complex modelling has traditionally been limited, and consider whether AI might result in illegal discrimination

This article is the second in a series investigating some of the economic consequences of using algorithms and the associated risks to businesses. Describing AI and its various forms,[1] our previous article highlighted that businesses face heightened risks when using AI, including regulatory and reputational risks.

As algorithms are used more and more in our daily lives, they raise concerns about their legality in areas where computer programs and complex modelling have traditionally played a limited role—for instance, regarding forms of discrimination that are often proscribed and carefully monitored under national law (e.g. when hiring new employees).

There are now numerous instances in which the outcomes of program-based processes are considered unfair or biased. As discussed in the previous article, one recent prominent example is the public uproar following the use of an algorithm by Ofqual (the Office of Qualifications and Examinations Regulation) in the UK to 'predict' students' A-level results this year when their exams were cancelled due to the COVID-19 pandemic.[2] Less overt cases include:[3] crime-prevention programs that target specific communities; online advertisement algorithms that are directed towards those who are less wealthy, offering pay-day loans; and job-recruitment programs that penalise women. Algorithms are also used to correct existing social imbalances.

How can AI be affected by human bias and be prone to being discriminatory? What went wrong with Ofqual's algorithm and how could this have been prevented? How can trusted and lawful algorithms be designed?

What is the link between AI and discrimination?

However sophisticated a program, algorithm or AI system may be, it is designed to find patterns in data (differences and similarities) that help it to accurately predict outcomes, match individuals, or classify objects or individuals.[4] By design, AI 'discriminates' as it separates datasets into clusters on the basis of shared characteristics, applying different rules to different clusters. This segmentation and the resulting differences in the rules applied simply reflect the best way that is identified by the algorithm to achieve a particular objective given the data.

Yet the programs involve human intervention at all stages: the algorithm and its objective function are coded by a programmer, the data used to train the algorithm is collected by a human, and the algorithm is rewarded for replicating, to some extent, human decisions. Human-driven discrimination, then, can be introduced at all stages of the process.

How can a program be affected by human biases?

As humans are involved at key stages of the development of an algorithm, programs can recreate, and reinforce, discrimination from human behaviour. Exactly how do human biases affect algorithms?

Individuals can be biased against others outside of their own social group, exhibiting prejudice, stereotyping and discrimination. Machines do not have all of these preconceptions and biases, but they are designed by humans. In addition, the way in which their programming operates may also lead to perverse feedback loops.

In practice, human biases can affect programs in a number of ways.

Unrepresentative or insufficient training and test data. If the training and test datasets are not representative of the overall population, the program will make incorrect predictions for the part of the population that is under-represented. However, even a representative sample of the population may not be enough: it needs to be sufficiently large such that the margin of error is sufficiently low across the various sub-groups of the dataset.[5]

The type of information provided to the program about the population. Although programs are designed to find patterns in the data, they operate based on the set of variables defined by the human programmer. If these variables are incomplete, approximate, or selected in a biased way (even unconsciously), the program will find patterns that may reflect the programmer's perceptions rather than providing an objective analysis.

As an example, if female applicants to a tech company have a significantly lower chance of being hired (based on historical data) than male applicants,[6] a CV-sifting program is likely to predict that female candidates are generally not as good as male candidates. By not considering the possibility that the hiring process may have been biased in the past, the design of the program repeats and automates human biases. A solution would be to exclude sensitive variables such as gender. However, such a simple fix may not work when other factors in the dataset are strongly correlated with the excluded sensitive variables, such as the type of extracurricular activities or courses taken.

This bias can also emerge when the algorithm uses one factor as a proxy for another because the information about the right factor is not known or included in the dataset. This may be the case, for example, when the program uses gender or race as a proxy for behaviour. In the case of car insurance, women in France used to pay lower premia than men, as they are, on average, less prone to accidents.[7] Ideally, the program would have taken into account characteristics that are associated with a lower probability of having an accident other than gender, but it is fundamentally interested only in parameters that enable it to predict the probability of accidents. If gender and the probability of accidents are correlated, discrimination can occur by generalising to an entire group something that is observed, on average, more frequently (but not always) within this group than across the entire population.

Prediction errors. The best program (from the perspective of its objective) is not perfect and still makes mistakes even with a representative dataset, as programs cannot model every particular set of circumstances. For instance, a credit-scoring algorithm can take into account only accessible factors such as employment history, available savings, and current debt. This means that some candidates who would be able to meet the financial obligations may still fail to obtain credit. This problem of 'collateral damage' can be exacerbated by poor design, by the programmer, of the model applied by the program.

The way the program learns, or does not learn. Machine learning programs that rely on reinforcement processes (i.e. learning from the impact of their previous actions) are subject to the equivalent of the human confirmation bias: a program may detect an action that encourages it to act further in the same direction, sometimes in a way that is socially discriminatory. A classic example is an algorithm used in the USA[8] that dispatches police to areas where more petty crimes were previously

reported. The initial decisions of where to send the police may be biased by racism. Though the original issue with this program was the 'tainted' data, the way the program operated could amplify the issue by reinforcing the bias.

Even programs that do not learn can exacerbate existing disadvantages in society. For example, algorithms used to optimise staffing in some professions (e.g. cafes),[9] although very effective at their objective of lowering costs and customer waiting time, have led to unpredictable working hours, as programs optimise workforce almost on a daily basis.

Designing algorithms that do not discriminate: lessons from economics

Economists have analysed individual biases that lead to discrimination and are familiar with the design of statistical models. As a consequence, economists are in a unique position to help design AI in a way that prevents creating or perpetuating unlawful discrimination. In fact, when undertaken properly, algorithm design may help to mitigate the impact of historical discrimination.

Establishing, and correcting, potential algorithmic discrimination can be done before the algorithm is implemented (ex ante), or once it has been rolled out (ex post), as illustrated in Figure 1.

Ex ante, the risk of implementing an unlawful algorithm can be reduced by carefully controlling the design of its objectives and the dataset used to train it. Controlling these objectives is likely to require **coordination** between the programming teams and those in charge of tackling discrimination. Training programmers to build 'ethical' AI is also a hot topic in the data science profession.[10]

More difficult to identify is bias arising due to the dataset failing to be **representative of the population**. In this regard, lessons could be identified from the design of statistical surveys, which are a common tool for decision-making and where, quite often, it is challenging to collect data in a way that is representative of the whole population.

When it is possible to influence data collection, a representative sample should be constructed. The first step would be to collect a **large enough random sample** of the population to ensure that all groups are sufficiently represented. It is also possible to use 'stratified sampling', in which the population is divided into sub-populations, and individuals are randomly selected within each sub-population. Each sub-group needs to be large enough to ensure reliable classifications or predictions.

If it is not possible to influence data collection, controls need to be introduced to ensure that the data is representative of the population. Simple summary statistics on the sample collected can be compared to the underlying population. If the sample is not representative, **more weight could be given to certain data points** to reflect the size of different sub-groups in the population.[11]

Finally, it is crucial to check that the list of factors that the algorithm uses is not biased. This is not an easy task. While humans can easily detect obvious potential sources of bias, they often cannot detect bias arising from **spurious correlations**. In such cases, ex post assessment could be used to identify any required changes to the original factors. Ex post, discrimination can be identified from the **results of the algorithm**, without any knowledge of its technical functioning. The idea is to analyse the impact of the program on various sub-groups within the population to check whether any is being treated differently from the others. This would be an extension of 'field experiments'. In a field experiment, two or more groups of individuals are randomly allocated to different situations and results are compared between the groups.[12] To avoid discrimination, field experiments have been used in **CV sifting**.[13] Fictitious and identical job applications are sent to companies, with the only difference being a criterion potentially prone to discrimination (for instance, a man versus a woman). As these applications are randomly sent to companies, it has been possible to identify whether certain groups are, on average, discriminated against.

This process could be extended to all types of algorithms. Newly programmed algorithms could be forced to make a large number of decisions on similar cases, except for one criterion subject to discrimination. The outcome of these decisions could be analysed to uncover potential discrimination before the algorithm is made public. This could have prevented the A-level controversy, described in the box overleaf.

Finally, the ex post assessment can include an analysis of whether the program **reinforces existing differences** within the population. By assessing which sub-group in the population may be adversely affected by the program's decisions, it is possible to assess whether the program tends to exacerbate existing imbalances.

It is conceivable that, having undertaken all checks, the program leads to part of the population being more adversely affected. In a way, that is a logical outcome of the program trying to discriminate between individuals to reach a particular objective. It would be contrary to insurers' business models, for example, if they did not take into account all the (legally) relevant parameters to decide insurance premia. Yet the outcome of this process is that some people end up paying significantly more for insurance.

If the results are statistically sound and non-discriminatory, but socially questionable, algorithms can be used to **proactively 'correct' the results** in a way that is socially preferable, especially when the algorithm is used by public entities (e.g. the justice system and schools). For example, in France, ParcourSup has the *explicit* goal of partially compensating for the lower likelihood of students from poorer backgrounds attending the most prestigious schools and universities. The system is designed such that colleges and universities use their own algorithms to select students. The ParcourSup algorithm then adjusts the universities' rankings in order to increase the proportion of students

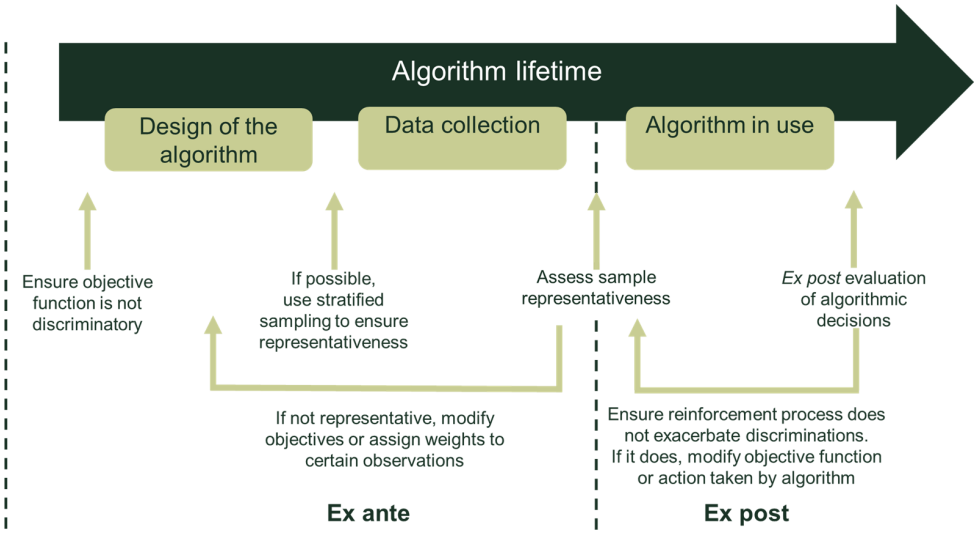

Figure 1 A framework to ensure discrimination-free AI

Note: The 'Algorithm in use' step does not mean that the algorithm is publicly available—it can also refer to an internal testing period.

Source: Oxera.

The 2020 A-level controversy in the UK

The algorithm used to predict this year's exam results in the UK induced differences in treatment across students. By imposing each school's 2019 distribution of grades in 2020, the algorithm failed to take into account how students' performance in a school may have differed from that of the previous year's cohort, disregarding predicted grades from teachers. This made it very unlikely that outstanding students in poorly performing schools would be allocated the top grades.[14] It also treated more harshly students from large schools taking popular courses than those in smaller schools or who took courses with fewer students.

To correct for this, the right technique to apply would depend on the source of the bias that led to the grade inflation in the first place. If the bias is perceived as applying equally across teachers (an application of the 'optimism bias'), a more socially acceptable approach could have been to downgrade teachers' grades by the same amount across the entire student population (e.g. by the average at the national level compared with the previous year). If the bias is at the school level (i.e. teachers in some schools are more affected by optimism bias than others), teachers' grades could have been downgraded by an amount specific to each school (e.g. by the average difference at each school compared with the previous year). In both approaches, the distribution of grades within schools is identified by teachers, and does not necessarily reflect the distribution observed in the previous year. It should be highlighted that the second approach assumes that a school could not achieve better results in 2020 than in 2019 overall—such an assumption could be considered unfair in itself.

It is important to identify the students who are more likely to be penalised by the application of the algorithm, especially if there are different ways to reach the same overall outcome. For example, Ofqual assessed how the 2019 grades would have turned out were its approach used in previous years (i.e. based on the 2018 rankings). The accuracy was found to be around 50–60%. Ofqual also argued that it considered distributional effects and concluded that 'the analyses show no evidence that this year's process of awarding grades has introduced bias'.

Source: Ofqual (2020), 'Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report'.

with scholarships that gain places at the prestigious institutions.

It is preferable to make this final adjustment as an explicit extra step in the algorithm design, as opposed to tweaking the initial algorithm to directly achieve the socially more acceptable results. Indeed, the final adjustment involves a degree of social judgement and perception, which must be undertaken carefully to avoid discrimination.

Reflections (in the mirror)

Although we use machines to make decisions, these decisions can be subject to human bias and discrimination. Since they use data collected by humans, are designed by humans, and have objectives driven by human interests, programs—regardless of their degree of sophistication—can create, reproduce or exacerbate discrimination. However, with the right approach to design and testing, it is usually possible to identify and reduce biases in programs.

Even when a program is designed and implemented in the correct way, it is still possible that its outcome is perceived to be unfair. This is because algorithms are a mirror for reality: we need to make sure that the mirror we are using is non-distortionary. However, if the underlying reality is disagreeable—for instance, as a consequence of historical and structural forms of discrimination—this will be reflected in a model's results. Nevertheless, after the necessary adjustments correcting for anti-discriminatory behaviour suggested in this article have been performed, a transparently designed algorithm can go one step further and correct outcomes that society deems unacceptable.

Contact

pascale.déchamps@oxera.com

Pascale Déchamps

ambroise.descamps@oxera.com

Ambroise Descamps

sarah.raviola@oxera.com

Sarah Raviola

gareth.shier@oxera.com

Gareth Shier

[1] See Oxera (2020), 'The risks of using algorithms in business: demystifying AI', September, https://bit.ly/3kEoFbv.

[2] For the full 319-page technical report from Ofqual on its algorithm design and outcomes, see Ofqual (2020), 'Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report', August, https://bit.ly/31PFNUd.

[3] For an overview of some of the programs with the highest potential for massive adverse societal impact, see O'Neil, C. (2016), *Weapons of Math Destruction*, Penguin.

[4] See Oxera (2020), 'The risks of using algorithms in business: demystifying AI', September, https://bit.ly/3kEoFbv.

[5] As an example, a dataset with 1,000 observations may be sufficiently large to get a prediction right 95% of the time. However, if the population (and the dataset) includes a relevant group that represents, say, 20% of the population (200 observations), it is possible that the algorithm will be right for this group only 20% of the time, instead of 95% of the time (with the prediction being wrong for the rest of the population only 1.25% of the time). This may be due to the sub-sample being too small to lead to reliable results across sub-groups.

[6] See Dastin, J. (2018), 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters*, October, https://reut.rs/34z312M.

[7] See, for example, LeLynx.fr, 'Pourquoi les femmes payent-elles moins cher?' (in French), https://bit.ly/2G5VnUa. It is now illegal in the EU to use gender as a rating variable for insurance premium calculation (Gender Directive of 2012).

[8] See Meliani, L. (2018), 'Machine Learning at PredPol: Risks, Biases, and Opportunities for Predictive Policing', *Assignment: RC TOM Challenge 2018*, Harvard Business School, November, https://hbs.me/2TyRyKc.

[9] See Quinyx.com (2020), 'Starbucks took a new approach', https://bit.ly/327jKc7.

[10] See Stolzoff, S. (2018), 'Are Universities Training Socially Minded Programmers?', June, https://bit.ly/2HGLTj2.

[11] For instance, if a sample contains fewer observations for women than for men, while in the population the groups are of equivalent size, a 'weighting' would duplicate certain observations with women to reach the proportions of the true population.

[12] This approach is commonly used in economics. For instance, the 2019 Nobel Prize was assigned to Banerjee, Duflo and Kremer 'for their experimental approach to alleviating global poverty'. See Bertrand, M. and Mullainathan, S. (2004), 'Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination', *American Economic Review*, **94**:4, pp. 991–1013, https://bit.ly/3kBJrJ1; and Ahmed, A. M. and Hammarstedt, M. (2008), 'Discrimination in the rental housing market: A field experiment on the Internet', *Journal of Urban Economics*, **64**:2, pp. 362–72, https://bit.ly/35BSB1K.

[13] See Bertrand, M. and Mullainathan, S. (2004), 'Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination', *American Economic Review*, **94**:4, pp. 991–1013, https://bit.ly/3oySxZh.

[14] For more discussions on how this algorithm was designed and implemented, see Thomson, D. (2020), 'A-Level results 2020: How have grades been calculated?', FFT Education Datalab, August, https://bit.ly/34z7UJ5; and Clarke, L. (2020), 'How the A-level results algorithm was fatally flawed', *New Statesman*, August, https://bit.ly/3mvxaWZ.