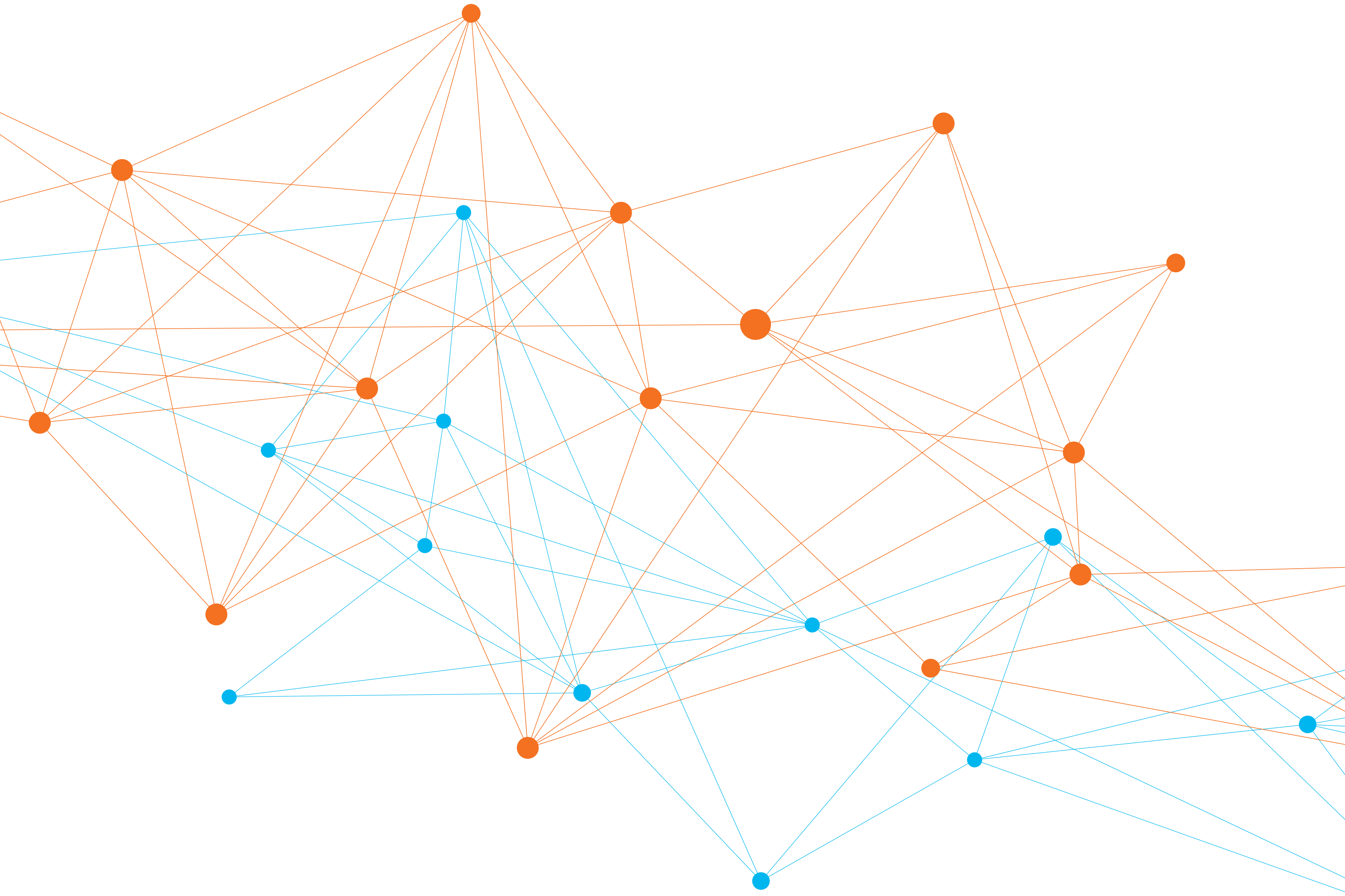
Institute *for the* Future of Work

Equality Task Force

Machine learning case studies



Putting people first

Our case studies explore how machine learning is being used in the workplace in twenty-first century Britain. The cases are hypothetical, involving fictional companies, decision-makers, and technical details, but they are based on existing research and real-world use cases. The case studies are designed to develop public understanding of how machine learning is used in the workplace, identify key questions and advance policy response in ways that would not otherwise be possible.

Published February 2020

Introduction

In recent years, we have seen a dramatic increase in the use of data-driven technology and systems to power decisions about the terms, conditions and nature of work. Within this transformation, new machine learning capabilities stand out.

Machine learning (ML) is a form of artificial intelligence (AI) which makes predictions and classifications based on the patterns and structures found within data sets. These case studies focus on supervised machine learning, in which a human specifies the outcome the model will learn to predict, the training data it will learn from, and the features the model will use to make the prediction. The machine learning model then learns which patterns and structures in the training data are statistically useful to predict the outcome the human has specified.

Machine learning is a socio-technical process, combining human judgement and statistical methods. Our case studies bring this to life by drawing attention to the role of humans within the process of machine learning, and the decisions machine learning is used to inform. The legal and ethical questions should focus on these human decision points, as they are critical to regulatory frameworks of discrimination and data protection.

The cases are structured around three key human decision-making points, which we explore in the three case studies below.

1

Outcome
What the ML model learns to do

The outcome is what the machine learning model is trained to predict. This outcome is always specified by a person. It can be a simple binary classifier: yes or no, guilty or innocent. It can be a ranking output of some kind, which places one prospective employee over another, or one piece of content over another. Or it can be a prediction: the probability an employee will remain in their post for longer than five years, or the probability a person will behave in a certain way in the future.

2

Training Data

What the ML model learns from

This is the data from which the model learns to predict the outcome. What makes machine learning powerful is the quantity and quality of training data the models learn from. The training data is selected and assembled by people. It contains information about the past, and so necessarily reflects existing patterns and the different opportunities and disadvantages different people have been afforded. This includes what kind of employee companies have tended to hire in the past; who performed well in a particular environment; or who got promoted.

3

Features

What the model uses to predict outcomes

Features are the variables a model uses to predict, rank, or classify the specified outcome. In the process of machine learning, models learn which features are statistically useful to predict some outcome and how to weigh them. A person can then change the particular weightings, or decide to include or exclude certain features, in order to improve the accuracy of the model or to respect laws and regulations.

Equality Law

The Equality Act 2010 makes it unlawful, in almost all contexts, to treat people differently in relation to work on the basis of protected characteristics like race, sex, religion, disability or sexual orientation (There are some limited exceptions in s159 Equality Act 2010 for proportionate steps to remediate disadvantage, except in relation to hiring). It also prohibits ‘indirect discrimination’, which is where a provision, criterion or practice (PCP) puts members of a group with a protected characteristic at a particular disadvantage which cannot be objectively justified. The causal link is between the characteristic and the disadvantage, not the use of the PCP and the disadvantage. If such a disadvantage is shown, the use of the PCP is unlawful unless it can be objectively justified by the person using it.

Data Protection

Data protection legislation contains additional limits on the use of sensitive personal data, and on solely automated decision-making. The precise application of these laws to the use of machine learning systems in the work place will be considered by the Task Force.

1

Hiring systems

Case study 1

Hiring systems

The Thor case study draws out key challenges in practice and accountability arising from the increasingly common use of ML systems in hiring. This is especially important because automated hiring systems will determine access to the labour market and opportunities for career development.

Founded in 1997, the telecommunications firm Thor took pride in its personalised hiring process. All candidates for senior positions were interviewed in person, whilst unsuccessful applicants for call centre jobs were offered individual feedback and the option of alerts for future positions. Thor had an excellent reputation for a fair hiring process and for being an equitable employer who treated its employees well. You were lucky if you got a job at Thor.

However, Thor's reputation for being a good employer had recently suffered. The press had accused Thor of unfairly disfavouring women and members of lower-income communities in its hiring process, as the proportion of women and people from C2DE backgrounds in the National Readership Survey (NRS) system of classification had fallen.[1] Worried about possible lawsuits, Thor commissioned a thorough review of how its hiring practices had changed over the last five years, led by a prominent public official. Thor decided the best way to confront its challenges was to be open about what was driving them, so as to hold themselves up as an example for other firms whose hiring practices were changing.

The problems began when Thor decided to change its hiring process in 2014. The productivity of Thor's call centre employees had diminished, with employees handling fewer calls per minute with lower average levels of customer satisfaction, whilst the turnover of senior executives had increased, with significantly more senior employees leaving in less than five years. Thor hired experts on data mining and machine learning to lead a team to develop three automated hiring systems: one for call centre employees; one for coders and software engineers, and one for executives.

Case study 1
Hiring systems

System 1

Call centre employees

The system for hiring call centre employees was developed by using data from the social media profiles of current employees, coupled with their employment records to predict the future job performance of candidates for positions at call centres.[2]

Training data

The system combined training data from two sources:

1. Data taken from the public social media profiles of existing employees at call centres across the UK Current call centre employees were asked to provide links to their social media profiles to complete Thor's annual employee satisfaction survey. Most those who completed the satisfaction survey provided links to their social media profiles (Thor had placed the box to tick if an employee did not wish to provide their social media profile at the bottom of the completed page in small text). This meant Thor had a large database of information on employees' public social media accounts (Facebook, Twitter, Instagram, etc.).[3]

2. Data about the job performance of existing employees. Thor had collected three years of performance data about its employees, which included information about the number of calls each employee handled per hour over time, and the average customer satisfaction rating they received for their calls.

These two data sets were combined to form the training data set for the machine learning model, which would form the backbone of Thor's call centre hiring system.

The system would combine the two scores, then rank the candidates according to their overall scores. The top 15% of candidates would automatically be invited to interview. About half of those invited to interview would be hired.

Outcome

Using a combination of these two data sets about existing employees, two machine learning models would be trained to predict two different outcomes for future employees. The first would predict the average number of calls an employee would be likely to handle per hour. The second would predict the average customer satisfaction rating they would receive for calls. These predictions would be for the first year of employment; employee turnover had increased so much that predictions beyond that were not especially relevant. The system would combine the two scores, then rank the candidates according to their overall scores. The top 15% of candidates would automatically be invited to interview. About half of those invited to interview would be hired.

Features

Thor decided their models could use any variable which appeared in the public social media profiles which employees had voluntarily shared, apart from any legally protected categories.

Thor was eager to maintain its reputation for excellent hiring and employment practices, so it made sure to test all its algorithmic systems. The results for its call centre models were unsettling: they found that about 60% of those recommended for interview were ABC1 in the NRS system of demographic classification, despite making up only 20% of applicants to call centres positions. Moreover, although 50% of applicants for call centre positions were female, only 30% of those recommended for interview were female.

Thor wanted to understand what was driving these results. Two factors were important. The first was that the social media profiles of employees and applicants from ABC1 backgrounds contained much less information than the social media profiles of more affluent employees and applicants. This meant both models were less accurate predictors of the specified outcomes for ABC1 applicants. Thor explored ways of acquiring additional data on a broader range of applicants, but found they were so expensive that they would undermine the cost-saving goal of automating the company's hiring practices.[4]

The second was that customer satisfaction ratings were higher on average for employees from middle- and upper-class backgrounds than lower-class backgrounds, and for male than for female employees. Thor hired a marketing firm to execute a qualitative study about why this was the case. They discovered that it had a lot to do with language. Most of Thor's customers were from ABC1 backgrounds, and customers from those backgrounds were more likely to give higher customer satisfaction scores to employees from similar backgrounds. Customers seemed to prefer to talk to people who sounded like them. Additionally, while Thor's customers were approximately 50% male and 50% female, both male and female customers gave higher customer satisfaction scores to male employees.

Thor explored ways of acquiring additional data on a broader range of applicants, but found they were so expensive they would undermine the cost-saving goal of automating the company's hiring practices.

Thor decided these were consumer preferences over which they had no control. Their job was to provide their customers with a good service, not question why customers preferred some employees over others.[5]

On the basis of this analysis, Thor decided to integrate the call centre hiring models into their hiring system. From 2015 to 2018, the proportion of ethnic minorities employed in Thor's call centres decreased by almost 20%, and the proportion of female employees decreased by about 15%.

System 2

Coders and software engineers

Training data

The system for hiring coders and software engineers used three types of training data. Firstly, Thor purchased Sales Navigator, LinkedIn's tool for recruiters, which gave them access to data about the LinkedIn profiles of past and prospective candidates.[6] Secondly, they used data scraped from the CVs submitted by successful applicants to Thor's tech team over the past five years. Thirdly, they used a database of information about the employees in Thor's tech team over the past five years, including information about the average length of time each employee stayed with the company and their Myers-Briggs personality type. This Myers-Briggs personality type was based on the standard four-dimension personality test used to assess characteristics such as extraversion, capacity to plan ahead, and different forms of reasoning.[7]

Outcome

The system for hiring engineers would be implemented in two stages.

The first was a simple classification task to predict whether applicants were 'suitable' or 'not suitable'. This classification would combine the results of three machine learning models, each of which predicted a different outcome. If any of these models produced a 'suitable' classification, the candidate would be placed on the shortlist for interview.

These three machine learning models were:

1. A model which determined whether the candidate had at least four years of experience in a similar coding company.

2. A model which determined whether the candidate had a computer science degree from a university similar to those where Thor's existing team had studied prior to joining.

3. A model which produced a numerical score from 1 to 10 of 'expressive skills' which determined how similarly the candidate expressed themselves in their CV to those of Thor's existing employees. Candidates with a score above 5 were classified as 'suitable'.

The second stage combined two predicted scores to produce a ranking of candidates, which would be used to determine the order in which they were interviewed. The first would predict the probability a candidate would remain in their position for more than five years, since Thor believed there were considerable costs in training and integrating employees with technical backgrounds. The second would predict the probability a candidate would be a particular Myers-Briggs personality type.

Following other employers, Thor's preliminary research showed that some personality types were more likely than others to be the most effective data scientists and engineers.

Following other employers, Thor's preliminary research showed that some personality types were more likely than others to be the most effective data scientists and engineers.[8] Thor would use data from LinkedIn, CVs, and the performance of past candidates to rank candidates shortlisted for interview by job tenure and personality type.

Features

The team used unsupervised natural language processing to devise clusters of terms found in the CVs and LinkedIn profiles of existing tech team employees in order to predict each of these two outcomes (job tenure and personality type) based on words used in CVs and LinkedIn profiles of candidates. A set of twenty variables were found to be meaningfully predictive: there was a strong interaction between these variable from within the training data sets, and the two outcomes. These variables were included in the model, but demographic traits like ethnicity, gender, and age were excluded.

Thor's testing revealed two problems. The first was that almost all candidates shortlisted for interview were men. Thor had deliberately set a low bar for calling a candidate to interview – such that a candidate had only to receive a 'suitable' classification on one of the three outcomes to be invited for interview. Nonetheless, this did not prevent the model from learning to use several features which correlated with gender to determine whether candidates would be classified as 'suitable' and invited to interview.

Case study 1
Hiring systems

Though one criterion for a 'suitable' classification was whether a candidate had a computer science degree, the universities the model recognised were based on the universities that existing Thor engineers had attended. Because these engineers were mostly men, the classifier was ignoring many other majority-female universities that applicants had attended, but that Thor's existing predominantly male engineers had not. Even more troubling, the model which gave candidates an expressive score tended to give weight to verbs like 'execute' and 'dominate', which were used more often by male candidates than female candidates. The underlying driver was that these were the kind of words Thor's existing predominantly male engineers had used in their applications.[9] The model simply learned to accurately predict the kind of candidate currently in Thor's tech team. What's more, since the team was small with only about 500 members, 50 of whom were women, the model simply did not have enough data to train an accurate model for women.[10]

The second problem was that even women who did make it to the interview shortlist were consistently being ranked below men. Two factors were driving this. First, the kind of personality traits Thor had found to correlate with high performance as an engineer or data scientist were found to be more common in men than women. The second was that the data in the training set showed that female engineers did not tend to stay as long in their jobs on average as male engineers, although the data did not distinguish between people's different reasons for leaving their positions: whether they were fired, hired by a competitor, or left to have children.[11]

The problem was that small differences between average rates of personality type and job tenure would become larger and larger over time, as the model was updated and retrained to reflect new data. Each increase in the gap might be imperceptible on its own, but the combined effect would be significant: Thor predicted it would hire half as many female engineers within ten years.[12]

Small differences between average rates of personality type and job tenure would become larger and larger over time, as the model was updated and retrained to reflect new data. Each increase in the gap might be imperceptible on its own, but the combined effect would be significant: Thor predicted it would hire half as many female engineers within ten years.

Case study 1
Hiring systems

Discussion

Proportionality

This case study draws attention to perhaps the most significant challenge machine learning presents: it reproduces the patterns of the past. Machine learning models are trained on data which encodes pre-existing patterns of opportunity and disadvantage, and identifies correlations from patterns of existing data without making judgments on causation. Without human intervention to decide what, if any, weighting should be given to particular pieces of information, the machines' predictions are likely reinforce those patterns. The Thor case study invites consideration of how a court might weight up the importance of an objective against the limitation of a protected right in this context.

Some key questions

- When is the Equality Act engaged? Could a 'PCP' be the choice of a training data set, decision-making model or an output?
- What degree of transparency about the purpose, model or approach taken is appropriate?
- How should we go about determining whether or not a choice is a 'proportionate' way of achieving a legitimate aim in the case?
- Can audits help Thor, or candidates applying for jobs, identify or defend potential cases of discrimination?

2

Job advertising

Case study 2

Job advertising

The Networkz case study explores the tension between use of social media by people in their private lives, and use of social media profiles by employers or prospective employees, by looking at the practice of targeted advertising.

Different impacts of ML at a systems, firm and individual level can be seen. The case draws out some key challenges in thinking about 'causation'.

Since 2010, Networkz has become one of the world's largest social media companies. Networkz aims to connect users and distribute content they would find engaging. Like other social media companies, Networkz's revenue comes from selling adverts distributed through its advertising delivery system, which used machine learning to predict which adverts users will click or comment on, like or share. Over the past nine years, adverts for jobs have become an increasingly important part of Networkz's revenue stream.

System 1

The simple version

Training data

There are two primary sources of data used in the system.

The first is data about users' general behaviour on the Networkz site. This included information the user has offered publicly, such as their post code, age, gender etc., as well as information about their online behaviour, such as what kind of content they have clicked on, liked, and shared, whom they tend to talk to and what kinds of emotions they tend to display in conversations with friends and family.

The second is specifically about users' engagement with job adverts. Job adverts have been classified using a range of factors including industry, income level, location, qualifications, and experience. Data about users and their online behaviour is then combined with this data about job adverts to identify correlations between certain characteristics (e.g. post code, gender) and behaviour (e.g. likes, contacts) with whether, in the past, users with those characteristics have clicked on certain types of jobs, to form a weighting score which predicts how likely a particular kind of user is to click on any particular job advertisement.

Outcomes

The first iteration of Networkz's advertising delivery system was reasonably simple. Two models were trained. The first would predict click probability: the probability that a user would click on a job advert. The second would predict application probability: given that a user had clicked on a job advert, the probability they would then apply for the advertised post. Users would then be shown the job adverts with the highest combined predicted score for clicks and applications.

Features

Though the outcomes these models predicted were reasonably simple, they used a huge range of features as inputs. This included everything from straightforward demographic information about age and gender, to information which the ML identified as having a correlation with likelihoods of 'clicking' and 'applying', but which the analysts could not explain as having an intuitive causal link with the likelihood of these outcomes, such as the average length of time a user stayed on a particular page, or the number of times per week they wrote a status and then deleted it. The models used over four thousand inputs in total, though this changed each time the model was retrained.

Over the course of six months, during which the models were fine-tuned and inputs were reweighted to reflect new data and further examples, Networkz's predictive models became extremely accurate in predicting who would be likely to click on a job advertisement and to apply for a job. For five years, the company refined their models, satisfied that the revenue from job advertisements went up year on year.

The model reinforced existing patterns of behaviour in the job adverts it offered to members of different groups.

However, an employee of Networkz decided to conduct some analyses of how the model was working for different social groups, focusing on the model performance and the average income of jobs shown for different social groups. They reached two conclusions. First, the model was performing better for women than men. It was more accurately predicting which jobs women would click on than which jobs men would click on. However, it was showing jobs with lower average income to women than to men. When the employee did some research, they found that women clicked on and applied for lower-income jobs than men.

When the employee examined the underlying data, she noted that there were far more female users in the training data, which helped to explain why the models were more accurately predicting their click and application behaviour. But she also noted that there were clear correlations between the types of jobs men and women tended to click on and apply to, which reflected prevailing social assumptions about 'women's' work and 'men's' work: women were more likely to click on care jobs, but men were more likely to click on other manual labour jobs. The jobs adverts that women tended to click on and apply to had a lower average income than the job adverts men tended to click on. Because women tended to click on these jobs traditionally filled by women, the model had learned to show these stereotypically 'female' and lower paid jobs to women.[13] In other words, the model reinforced existing patterns of behaviour in the job adverts it offered to members of different groups.

The employee discussed this with Networkz's leadership. They decided that Networkz had a responsibility to ensure they did not exacerbate existing patterns in user behaviour, but they did not have a responsibility to change those patterns in behaviour. It was not their job to decide which kinds of jobs their users wanted to click on or apply to, nor to set the rates of pay attached to them. However, they decided to change the model: they would add in two more factors and set a maximum gap between the average incomes offered to men and women.[14]

System 2

The complex version

Training data

The training data would be the same as in the simpler versions of the model, but two more pieces of information would be added. First, Networkz would ask employers to state whether the application they received was successful or not. Second, Networkz would ask employers to state, on a scale of 1 to 10, how satisfied they were with the employee.

Outcomes

Two further models would then be added. In addition to predicting the probability that someone would click on a job advert and apply to the job if they did, models would also be trained to predict, once they had applied, the probability their application would succeed and the likelihood and employer would be satisfied if they did succeed.

The idea was that if it were the case that women and men wanted certain kinds of jobs, and had the aptitudes to get them when they applied, then those candidates would be successful in applying to them and would receive high employer satisfaction scores. This would tease out whether it was really a form of bias in Networkz's delivery system for job adverts, or whether it was simply a reflection of the existing skewed preferences of users and employers as to who wanted certain types of work; who employers wanted to employ, and who they found good at the work when they did.

Networkz also decided to set a maximum gap between the average wage levels for the job adverts shown to men and women of 5%. In other words, Networkz's models would ensure that the average income of job adverts would not be any more than 5% higher for men than women.

> The plan did not quite work as they had expected. The additional features actually increased the gap in the average incomes of job adverts shown to women and men.

The plan did not quite work as they had expected. The additional features actually increased the gap in the average incomes of job adverts shown to women and men. In other words, the patterns of application success and job satisfaction were even more gendered than the patterns of click and application probability. But this meant the maximum 5% income gap was forcing the system to show women and men adverts for jobs they were less likely to click on; less likely to apply for; less likely to get if they did apply, and in which employers were less likely to report satisfaction if they did apply.

Networkz were left with a conundrum. They could remove the 5% maximum income gap, on the basis that the unequal outcomes their system produced were simply the result of unequal patterns in the user behaviour. Networkz's service was to match users with jobs they wanted to and could successfully apply to, and to match employers with candidates they will be happy with, whereas the structure of society was not Networkz's responsibility.[15] Or, they could keep the 5% income gap, immunise themselves from possible criticism for disparate outcomes and indirect discrimination suits, but risk dissatisfied customers.

Discussion

Causality

This case study raises two difficult questions for the pursuit of equality in the age of AI. Both are about causality.

Machine learning is a world of correlation, not causation. Machine learning models use whichever variables they find to be statistically useful to predict some outcome. They have no predispositions or inclinations to use race or gender in a decision-making system. But if race and gender are correlated with statistically-useful outcomes, the model will reproduce racial and gender inequalities without race and gender playing any causal role in the decision.

Some key questions

- Does the traditional notion of causation make sense when applied to automated decisions involving machine learning and based on correlations? Do we need to understand why correlations are being drawn?
- Given new capabilities, is it possible to isolate one reason or ‘protected characteristic’ for the treatment of candidates? Do existing protected characteristics adequately cover the criteria considered?
- Should Networkz be required to make adjustments to its model to find a less discriminatory means of achieving the desired outcome?
- What level of recording, disclosure and access to information is required to make sense of causation in this context?

3

Talent management

Case study 3

Talent management

The Brill case study draws out pressing challenges arising from automated employee management, monitoring and enforcing productivity in order to save costs. IFOW research has found that this practice is commonplace in the service sectors, even in traditional workspaces like the shop floor.

Brill is the largest clothing retail brand in the UK Brill has led the market in online sales, and now over half of its business is online. It has sold many of its high-street stores and bought warehouses in suburban areas to process the shipment of goods across the country.

Five years ago, Brill rolled out its 'XFN' system. XFN is a comprehensive system for, as Brill puts it, 'giving our employees the best experience and getting the best out of our employees'. The XFN program tracks and monitors how the speed at which employees select and move goods, how much time they spend on email and workplace chat forums, as well as their performance and promotions over time.

Training data

To ensure that XFN had access to the broadest range of information, Brill began collecting as much data about employees as possible. This began by aggregating data about their behaviour on Brill's workplace systems, such as email and workplace chat. Brill then used information from workers' publicly available social media accounts (such as their Facebook and Twitter pages) which they combined with their workplace data, to see if particular pieces of information seemed to correlate with faster and/or more accurate performance.

Brill also used their physical spaces to gather data. They installed small cameras around offices, retail stores, and warehouses. Employees were told that these cameras were in place, and would help with product and workplace management.

Office cameras employed facial recognition software to identify employees, to track their movements, as well as, where possible, their interactions and emotional signals (such as smiling or frowning). This information was also

Case study 3
Talent management

combined with data from Brill's internal digital systems, to see what correlated with faster and/or more accurate performance.

Retail store cameras and analysis of online ordering patterns gathered data to detect daily and annual patterns in the demands of customers for different products, based on tracking customer movements and empty spaces on shelves in stores, as well as the patterns in the movements of workers over time.[16]

The final step was to introduce wearable wrist bands for Brill warehouse employees.[17] Brill explained that the wrist bands would promote health and safety, through monitoring employees' physical positions as they lifted and moved goods. As well as this, however, the wrist bands also vibrated to alert employees to messages telling them to go towards the areas in which they were most needed, or to alert them if their productivity dropped. The wrist bands also sent the data back to XFN, where it was processed and combined with data from Brill's own digital systems, data taken from social media accounts and online ordering, and cameras.

Outcome

Over time, Brill came to use XFN for many employee monitoring and management decisions. There were three particular functions for which XFN was used:

Performance

XFN was used to produce a weekly employee performance report, which was sent to employees and their line managers. These reports would measure each employee's performance on a range of metrics; not only against what was expected of an average employee, but against an individual employee's predicted performance, based on past data about their performance over time.

For warehouse workers, XFN reported the number of hours worked and goods moved; for store-floor workers, their sales and customer satisfaction ratings were reported.

If employee performance had sufficiently improved to produce a suggested promotion, XFN would predict the level of pay rise likely to minimise the chance of an employee leaving, whilst also minimising costs.

Recommendations

These weekly reports were then aggregated into reports produced bi-annually, which made recommendations about pay rises, disciplinary actions, or promotions.

These recommendations were made not only on the basis of data about actual past performance, but also on the basis of predicted future performance. Future performance was predicted using the correlations between data about past performance and other information the system had found to be statistically useful in predicting future performance, including time spent on workplace chat, facial expressions on workplace cameras, and speed of moving around warehouses.

Over time, XFN learned how to predict the probability that an employee might leave if they were not given a pay rise or promotion. This prediction combined information about promotions and pay rises with average job tenure of past employees, along with other information from employees' workplace accounts. If employee performance had sufficiently improved to produce a suggested promotion, XFN would predict the level of pay rise likely to minimise the chance of an employee leaving, whilst also minimising costs.

Allocation

Finally, XFN was used to allocate employees. Data about customer demand was used to create weekly shifts for employees in warehouses and retail stores. XFN also offered predictions about which employees would be most willing to accept relocation, or be willing to having their hours changed, in order to effectively allocate workers.

The decisions based on XFN data correlation appeared to work as intended. Wage costs reduced significantly, and average worker productivity increased. Staff surveys initially suggested that employee satisfaction increased after the introduction of XFN technology, but this soon eroded as Brill's employees grew increasingly frustrated with the system. There were several reasons for this frustration.

Female, and Black, Asian, and minority ethnic (BAME) employees began to find that their shift patterns would change without warning in ways which had not happened before the introduction of XFN, and which (Brill discovered when it measured this) did not happen so often to male and/or white employees.
Male employees were offered unexpected pay rises but encouraged to change locations again in ways which did not happen to women, and had not happened before the introduction of XFN. The system even seemed to confuse senior BAME employees for warehouse workers. Workers were unhappy when they realised that XFN's cameras were being used not just for security, but to track workers' movements.

Nonetheless, Brill decided to continue to use XFN – it would be too costly to remove XFN from Brill's employee management and monitoring system. Instead, Brill offered mechanisms for employees to offer feedback on XFN and to dispute its recommendations or reports.
They also pledged to consult an 'XFN Rep' board of workers whenever further changes to the XFN system were to be rolled out.

> Female, and Black, Asian, and minority ethnic employees began to find that their shift patterns would change without warning in ways which had not happened before the introduction of XFN, and which (Brill discovered when it measured this) did not happen so often to male and/or white employees.

Discussion

Accountability

The final case study brings out the complex and connected ways in which machine learning can be used to monitor and manage employees. It draws attention to a familiar structural problem in the workplace which is exacerbated by machine learning. The machine learning in the workplace increases the asymmetry of information between employers and employees, or prospective employees throughout the stages of employment.

Machine learning also encourages employers to shift from focusing on performance to prediction. Instead of evaluating what employees can do, the proliferation of data encourages employers to try to get a picture of who employees are, then use that to predict what they might do in the future. The use of data to predict individual and collective future actions may encourage decision-making based on comparisons between similar individuals and groups.

Some key questions

- What requirements for consent, and other current protection, applies to the cumulative gathering and use of data for different purposes at work? Can consent ever be freely given at work? Can data gathered for one purpose be used for another?

- What does this mean for the accountability of different players and human decision-makers?

- What are the implications for the employer-employee relationship of moving towards predictive models?

- Are there any gaps in equality law or in enforcing equality law? Are other forms of accountability needed to promote equality in the age of AI?

> “Correlation isn’t causation. These algorithms aren’t social scientists or human resource specialists.
>
> Equality Task Force member

Endnotes

[1] Ipsos MediaCT, “Social Grade: A Classification Tool” (Ipsos, 2009), https://web.archive.org/web/20160315075958/https://www.ipsos-mori.com/DownloadPublication/1285_MediaCT_thoughtpiece_Social_Grade_July09_V3_WEB.pdf.

[2] Marie-Andrée Weiss, “The Use of Social Media Sites Data By Business Organizations in Their Relationship with Employees,” *Journal of Internet Law* 15, no. 2 (2011): 16–27.

[3] Gordon B. Schmidt and Richard N. Landers, *Social Media in Employee Selection and Recruitment: Theory, Practice, and Current Challenges* (Cham: Springer International Publishing, 2016).

[4] The problem of unrepresentative data. Companies often have less interest in reaching those at the bottom of the income distribution, and so, there is less available data on them. Solon Barocas and Andrew D. Selbst, “Big Data’s Disparate Impact,” *California Law Review* 104, no. 3 (June 1, 2016): 671–732.

[5] Jonathan Guryan and Kerwin Kofi Charles, “Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to Its Roots,” *Economic Journal* 123, no. 572 (2013): F417–F432.

[6] George Anders, “LinkedIn’s Mountain Of Data Is Now A $1,200 Sales Tool,” Forbes, accessed September 30, 2019, https://www.forbes.com/sites/georgeanders/2014/07/31/linkedins-mountain-of-data-is-now-a-1200-sales-tool/.

[7] Adrian Furnham and Paul Stringfield, “Personality and Work Performance: Myers-Briggs Type Indicator Correlates of Managerial Performance in Two Cultures,” *Personality and Individual Differences* 14, no. 1 (1993): 145–153, https://doi.org/10.1016/0191-8869(93)90184-5.

[8] Murad Ahmed, “Is Myers-Briggs up to the Job?,” Financial Times, February 11, 2016, https://www.ft.com/content/8790ef0a-d040-11e5-831d-09f7778e7377.

[9] Caroline Criado Perez, *Invisible Women: Data Bias in a World Designed for Men* (Harry N Abrahms, 2019).

[10] Word2Vec, among others, has been found to encode sexist representations of language inherent in the data it is trained on. https://www.technologyreview.com/s/602025/how-vector-space-mathematics-reveals-the-hidden-sexism-in-language/.

[11] Jahna Otterbacher, Jo Bates, and Paul Clough, “Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results,” vol. 2017-, CHI ’17 (ACM, 2017), 6620–6631.

[12] Jordan Weissmann, “Amazon Created a Hiring Tool Using AI. It Immediately Started Discriminating Against Women.,” Slate Magazine, October 10, 2018, https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html.

[13] Linda Miller and Rowena Hayward, “New Jobs, Old Occupational Stereotypes: Gender and Jobs in the New Economy,” *Journal of Education and Work* 19, no. 1 (2006): 67–93.

[14] The structure of this example mirrors the argument the U.S. Department of Housing and Urban Development is making about Facebook’s advertising delivery system in an ongoing discrimination suit. Department of Housing and Urban Development, “Charge of Discrimination,” 2019, https://www.hud.gov/press/press_releases_media_advisories/HUD_No_19_035.

[15] This raises the question of the scope and reach of indirect discrimination law. For an overview of indirect discrimination, and how it relates to concepts of fairness. Hugh Collins and Tarunabh Khaitan, *Foundations of Indirect Discrimination Law* (Oxford; Hart Publishing, 2018).

[16] https://www.telegraph.co.uk/news/2019/09/27/ai-facial-recognition-used-first-time-job-interviews-uk-find/.

[17] Ceylan Yeginsu, “If Workers Slack Off, the Wristband Will Know. (And Amazon Has a Patent for It.),” The New York Times, February 1, 2018, sec. Technology, https://www.nytimes.com/2018/02/01/technology/amazon-wristband-tracking-privacy.html.

Bibliography

Abram, Morris.
"Affirmative Action: Fair Shakers and Social Engineers." *Harvard Law Review* 99, no. 6 (1986): 1312.

Abrams, David, and Chris Rohlfs.
"Optimal Bail and the Value of Freedom: Evidence from the Philadelphia Bail Experiment."
Economic Inquiry 49, no. 3 (2011): 750–770.

Anders, George.
"LinkedIn's Mountain Of Data Is Now A $1,200 Sales Tool." Forbes. Accessed September 30, 2019.
https://www.forbes.com/sites/georgeanders/2014/07/31/linkedins-mountain-of-data-is-now-a-1200-sales-tool/.

Alexander, Michelle.
The New Jim Crow: Mass Incarceration in the Age of Colorblindness.
Revised edition. New York: New Press, 2011.

Bagenstos, Samuel R.
"Implicit Bias' Failure." *Berkeley Journal of Employment and Labor Law* 39, no. 1 (2018): 51.

Bagenstos, Samuel R.
"The Structural Turn and the Limits of Antidiscrimination Law." *California Law Review* 94, no. 1 (2006): 1–47.

Barocas, Solon, and Andrew D. Selbst.
"Big Data's Disparate Impact." *California Law Review* 104, no. 3 (June 1, 2016): 671–732.

Calders, Toon, and Indrė Žliobaitė.
"Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures."
In *Discrimination and Privacy in the Information Society*, 43–57. Studies in Applied Philosophy, Epistemology and Rational Ethics. Berlin: Springer, 2013.

Cass R. Sunstein.
"The Anticaste Principle." *Michigan Law Review* 92 (August 1, 1994): 2410–2649.

Chouldechova, Alexandra.
"Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments."
FATML 2016 Conference Paper, 2016.

Collins, Hugh, and Tarunabh Khaitan.
Foundations of Indirect Discrimination Law. Oxford: Hart Publishing, 2018.

Criado Perez, Caroline.
Invisible Women: Data Bias in a World Designed for Men. Harry N Abrahms, 2019.

Department of Housing and Urban Development.
"Charge of Discrimination," 2019.
https://www.hud.gov/press/press_releases_media_advisories/HUD_No_19_035.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang.
"The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges †." *American Economic Review* 108, no. 2 (2018): 201–240.

Dobbie, Will, and Crystal S. Yang.
"Equal Protection under Algorithms: A New Statistical and Legal Framework." *Working Paper*, 2019.

Guryan, Jonathan, and Kerwin Kofi Charles.
"Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to Its Roots."
Economic Journal 123, no. 572 (2013): F417–F432.

Hurley, Mikella, and Julius Adebayo.
"Credit Scoring in the Era of Big Data." *Yale Journal of Law & Technology* 18, no. 2 (2016): 148.

Jeff Larson, Julia Angwin.
"How We Analyzed the COMPAS Recidivism Algorithm." ProPublica, May 23, 2016.
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Jolls, Christine.
"Antidiscrimination and Accommodation." *Harvard Law Review* 115, no. 2 (2001): 642–699.

Bibliography

Khaitan, Tarunabh.
A Theory of Discrimination Law. Oxford: Oxford University Press, 2015.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein.
“Discrimination in the Age of Algorithms,” 2019.
http://arxiv.org/abs/1902.03731.

Kleinberg, Jon, and Sendhil Mullainathan.
“Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability,”
September 12, 2018.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan.
“Inherent Trade-Offs in the Fair Determination of Risk Scores.” *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017.

Moreau, Sophia, and Deborah Hellman.
Philosophical Foundations of Discrimination Law. Oxford: Oxford University Press, 2013.

Miller, Linda, and Rowena Hayward.
“New Jobs, Old Occupational Stereotypes: Gender and Jobs in the New Economy.”
Journal of Education and Work 19, no. 1 (2006): 67–93.

Otterbacher, Jahna, Jo Bates, and Paul Clough.
“Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results,”
2017-:6620–6631. CHI ’17. ACM, 2017.

O’Donnell, Renata M.
“Challenging Racist Predictive Policing Algorithms Under the Equal Protection Clause.”
New York University Law Review 94, no. 3 (2019): 544–580.

Peck, Don.
“They’re Watching You at Work – The Atlantic.” The Atlantic, December 2013.
https://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/.

Schmidt, Gordon B., and Richard N. Landers.
Social Media in Employee Selection and Recruitment: Theory, Practice, and Current Challenges.
Cham: Springer International Publishing, 2016.

Sullivan, Charles A.
“Employing AI.” *Villanova Law Review 63*, no. 3 (2018): 429.

Sunstein, Cass R.
“Why Markets Won’t Stop Discrimination.” *Social Philosophy & Policy* 8, no. 2 (1991): 22–37.

Sunstein, Cass R.
“Algorithms, Correcting Biases.” *Social Research: An International Quarterly* 86, no. 2 (2019): 499–511.

Tobia, Kevin.
“Disparate Statistics.” *The Yale Law Journal* 126, no. 8 (2017): 2382–2420.

Weiss, Marie-Andrée.
“The Use of Social Media Sites Data By Buisiness Organizations in Their Relationship with Employees.”
Journal of Internet Law 15, no. 2 (2011): 16–27.

Weissmann, Jordan.
“Amazon Created a Hiring Tool Using AI. It Immediately Started Discriminating Against Women.”
Slate Magazine, October 10, 2018.
https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html.

Yeginsu, Ceylan.
“If Workers Slack Off, the Wristband Will Know. (And Amazon Has a Patent for It.).”
The New York Times, February 1, 2018, sec. Technology.
https://www.nytimes.com/2018/02/01/technology/amazon-wristband-tracking-privacy.html.

Authors

These case studies were designed and written by Josh Simons, with input from Logan Graham and Anna Thomas.

Equality Task Force

Members are:

Helen Mountfield QC
Principal of Mansfield College,
University of Oxford (Chair)

Alasdair MacDonald
Director of Policy
Equality and Human Rights Commission

Professor Helen Margetts OBE
Professor of Internet and Society
Oxford Internet Institute

Dr Anne-Marie Imafidon MBE
Founder of Stemettes
Trustee of IFOW

Caroline Stroud
Partner
Freshfields Bruckhaus Deringer

Dr Reuben Binns
Research Fellow in Artificial Intelligence
Information Commissioner’s Office

Sue Ferns
Chair of Unions 21
Deputy Secretary of Prospect Union

Edward Houghton
Head of Research, CIPD

Josh Simons
Research Fellow, IFOW

Logan Graham
Research Fellow, IFOW

Anna Thomas
Director of IFOW

The Equality Task Force is generously supported by the international law firm, Freshfields Bruckhaus Deringer.

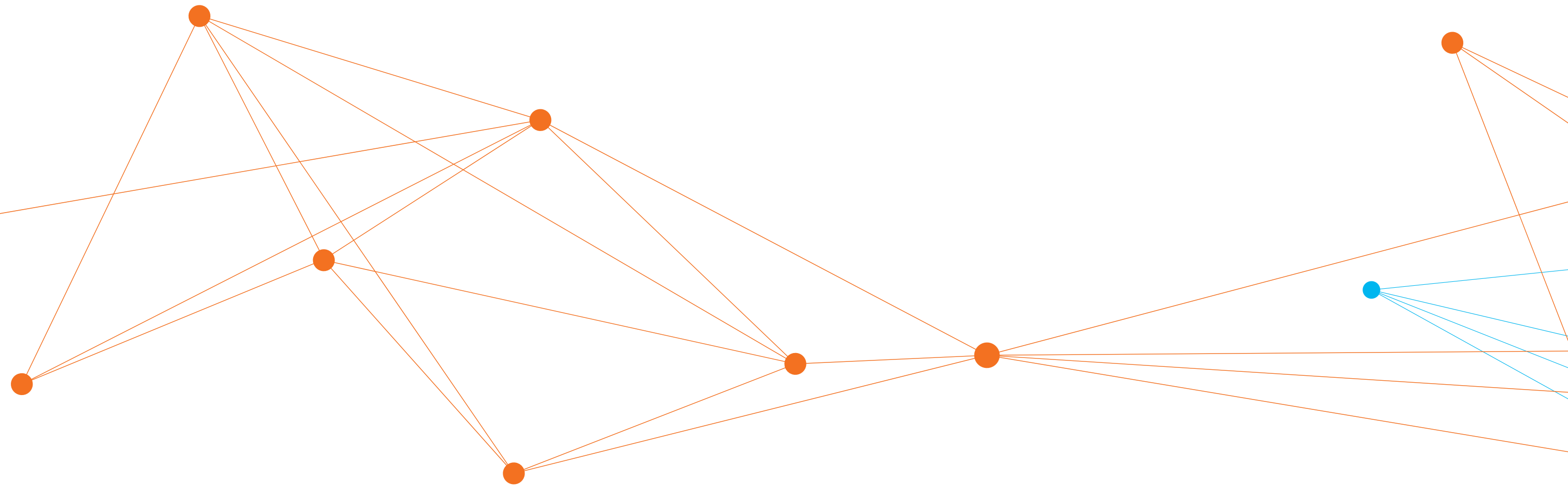
IFOW is grateful to all members of the Task Force for their time and contributions.

> “If an employer can’t explain why they recruited someone, isn’t that like saying: ‘they’re just my kind of person’?
>
> Equality Task Force member

Institute *for the*
Future of Work

Somerset House, Strand
London WC2R 1LA
T +44 (0)20 3701 7633

www.ifow.org
@_futureofwork



Registered charity number 1180718