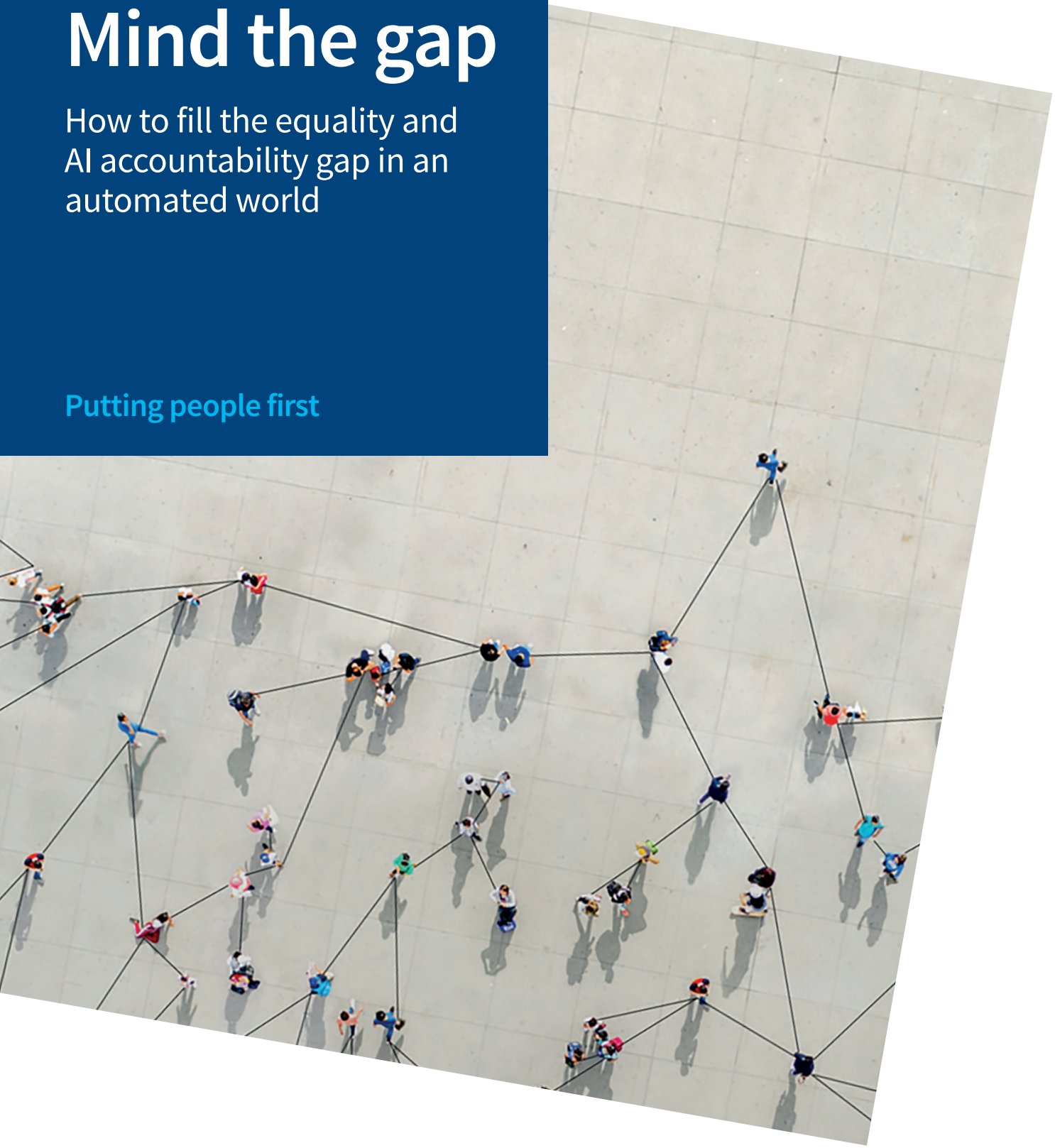
Institute for the Future of Work

Equality Task Force

Mind the gap

How to fill the equality and AI accountability gap in an automated world

Putting people first



“During the epidemic we have been made more powerfully aware of entrenched inequalities across the globe… there can be no doubt that they reflect structural inequality in our society which has to be addressed.

Rt Hon Michael Gove, Ditchley Park Lecture (June 2020)[1]

Published
October 2020

Executive summary

In September 2019 the Institute for the Future of Work established a cross-disciplinary Equality Task Force (ETF) to examine how algorithms and artificial intelligence impact equality and fairness at work.[2]

The ETF was chaired by Helen Mountfield QC and included academics (specialising in machine learning, internet, law), regulators (EHRC, ICO) unions (Prospect), trade bodies (CIPD), and business (Freshfields). It conducted its work between November 2019 and October 2020. This report sets out the ETF's findings and recommendations.[3]

Our starting point is that data, and data processing, are not neutral. Technology and algorithms do not 'make decisions': they process and use data that was pre-selected by human beings in ways devised by human beings, for purposes determined by human beings.

Human choices about how to build and use data-driven technologies are never neutral. Nor are the outputs of these technologies 'objective' or impartial. The information that is selected for analysis by data-driven technologies reflects both our assumptions about and the realities of the past. Data-driven technologies accurately capture the undesirable outcomes of the past and project them forwards, into the future.

In simple terms, all other things being equal, an algorithm based on historic data will assume that tomorrow's employees of a particular race, class, family background, educational background, gender and location will replicate the historic performance and results achieved by 'people like them'. Data-driven technologies offer unique and powerful opportunities to businesses to meet new challenges and understand patterns of behaviour and treatment. But, without thoughtful and careful intervention, they offer unsound and profoundly anti-aspirational bases for decision making.

Executive summary

Part 1

The myth of neutrality

Part 1 celebrates the potential and opportunities afforded by artificially intelligent ('AI') systems. It also explains why we must take the role of human agency far more seriously. Human beings select the input data, design the rules by which the data is processed and rely on the outputs. It is these decisions that shape the experiences and outcomes for workers as access to work, and the full range of employer functions, are digitised. Designers, coders and analysts, as well as employers, can and should be accountable for the ways in which they exercise their human agency in the design and deployment of AI systems.

Part 2

Structuring accountability

Part 2 examines how the operation, language and culture of data-driven technologies make decision-making both more diffuse and less transparent. Relatively few people understand the terminology or internal workings of AI, machine learning and mass data processing. This means that human agency is often obscured. It also means that those who make and control the technology, and data that feeds it, exercise the real power. They take decisions that will profoundly affect the lives of others on a daily basis. But the 'invisible' nature of what they do makes it more difficult to hold them accountable.

We argue that they should be accountable. Those at the heart of the data economy should consciously examine the adverse impacts of their work, especially equality impacts. That is both the burden and the privilege that comes with exercising agency on behalf of millions of other individuals. Meaningful accountability will help data-driven technologies serve the public interest, rather than becoming vehicles that reinforce unconscious biases and entrench inequality. Human agency must be affirmed, not removed.

Part 3 and 4

The regulatory ecosystem and key challenges and gaps

Part 3 and 4 analyse the existing legal and regulatory frameworks that are relevant in this area. We conclude that there is a regulatory lacuna in relation to data-driven technology. The Equality Act 2010 has much to offer but it was not designed with AI or machine learning systems in mind. The existing regulators lack effective levers to hold big data companies to account when they take decisions that have unfair and discriminatory results. These parts of the report are built around a series of case studies that illuminate both the strengths and weaknesses of the existing framework of law. These case studies demonstrate that more can and should be done.

Part 5

A new path forward: An Accountability for Algorithms Act

Part 5 sets out our central proposal: an overarching Accountability for Algorithms Act. That Act will direct and inform policy, standards and behaviours. It will impose new duties on mass data companies both at the *ex ante* stage – when designing AI and ML systems – and when determining how to use the outputs of those systems. It will change the behaviours and ethos of data-driven technology companies. They will be required to recognise that both they and the algorithmic systems they deploy are not neutral; their actions and their technologies can reinforce and perpetuate inequality. However, they can also be powerful engines for change. The Accountability for Algorithms Act will help those companies to become agents for the promotion of equality and fairness.

Although the ETF has focused on the use of algorithms, machine learning (ML) and Artificial Intelligence (AI) at work, we also highlight the potential wider significance of our recommendations for regulation of AI in other spheres.

To promote innovation and public good, and to ensure that the collective harms caused by statistical tools replicating past patterns of social inequality are not projected, unaccountably, into the future, the ethico-legal principle of equality between citizens and social groups has to be a central pillar of our societal and regulatory response.

Governance and regulation

We need a new approach to governance and regulation of data-driven, machine-based decision making. This approach must be principle-driven and human-centred, work across the entire innovation cycle, shift our emphasis to preventative action, and align our legal regimes and regulators. We need to articulate the objectives we as a society want technology to serve, to ensure that we govern the design and deployment of technology in ways which advance these purposes. This demands a review of existing laws and governance structures, to make sure that they address and advance these underlying purposes in a changing world, rather than straining existing laws and regulatory structures to 'fit' new technologies.

Innovation and public good

To promote innovation and public good, and to ensure that the collective harms caused by statistical tools replicating past patterns of social inequality are not projected, unaccountably, into the future, the ethico-legal principle of equality between citizens and social groups has to be a central pillar of our societal and regulatory response. We think this is necessary to ensure that data-driven technologies are built and used in the public interest. This is why there needs to be sharp focus on equality in the Act. But the other more established principles in AI governance need a statutory footing too.

Accountability for Algorithms Act

We propose a fresh approach to the regulation and accountability of data-driven systems, including AI and machine learning: a new Accountability for Algorithms Act. This is the cleanest and most pragmatic way to ensure that the specific allocation of responsibility and actions needed are clearly understood, undertaken and effectively enforced. And it will offer the direction that actors across the technology cycle – and the public – are demanding.

In Part 5 to this report, we outline a first framework for the Act. This is intended as a framework for wider consultation, development and for the addition of parts which fall outside the scope of the ETF, and those which require sector-specific attention.

> The digitisation of life is overwhelming the boundaries of conventional legal categories, through the volume of information which is gathered and deployed and the speed and impersonality of decision-making which it fosters. The sense is of a flood in which the flow of water moves around obstacles and renders them meaningless. Law needs to find suitable concepts and practical ways to structure this world in order to reaffirm human agency at the individual level and at the collective democratic level…

Lord Sales, Sir Henry Brooke/BAILII lecture on "Algorithms, Artificial Intelligence and the Law", Freshfields (Nov 2019)

Introduction

We are at a pivotal moment in the evolving relationship between data-driven technology, policy-makers and society. Data-driven technologies, including artificial intelligence and machine learning, are relatively new and extraordinarily powerful tools in the hands of companies, governments and society as a whole.

The advent of mass data processing, AI and ML is an exciting moment in human history. Data-driven technologies offer previously unimaginable opportunities to us all. Our core concern is that these technologies should be used in the public interest, particularly in the workplace.

Until recently, this was an issue which received little public attention. But in August 2020 the assignment of A-level grades by the Ofqual algorithm, which based its predictions on past performance of specific schools, thrust algorithmic accountability into the spotlight. The inherent dangers of uncritically using data about the past to make predictions about the future were revealed. While there were many differing views as to the solution, perhaps the most important product of the 'Ofqual crisis' was the broad consensus that the outcomes of the algorithm were unfair. This report seeks to build on and add analytical insight to that consensus.

There was widespread recognition in the summer of 2020 that student grades should not be assigned or weighted on the basis of past school performance. A mass data approach – at least in this case – collapsed individual students into the aggregate of their personal characteristics and background. As a result, their range of possible outcomes was informed by the results achieved by previous generations 'like them.' The data inserted into the algorithm reflected the systemic inequalities of the past. The algorithm then projected those inequalities into the future. That result was deeply troubling to many in the United Kingdom.

Whatever the merits of the particular statistical model Ofqual built, the case has put a spotlight on the need for robust mechanisms for accountability in the design and deployment of data-driven technologies.

Invisible data-driven technologies involving mass data processing are transforming work across the country. This transformation is accelerating[4] through the pandemic, which has exposed the structural inequalities in access, terms and quality of work. The reconvened Future of Work Commission[5] has pinpointed the acceleration and pervasive use and impacts of data-driven technologies, fed by increasingly detailed and complex data sources. While holding huge potential to increase efficiency, augment human capabilities and remove drudgery, benefits and adverse impacts are not spread evenly across demographic groups, occupations or places.[6] As the scale and speed at which these tools are adopted, so must the pace, breadth and boldness of our policy response.[7]

Our methodology

To understand and identify the issues at stake, the ETF used three case studies[8] to help explore increasingly common uses of one class of data-driven technology, machine learning, in three main areas of work: hiring, management and performance review, and how it affects the experiences and needs of working people. We have focused on machine learning ('ML') because ML technologies sharpen and 'supercharge' the impacts and challenges we examine.

Our case studies are hypothetical, involving fictional companies and decisions, but reflect ways in which ML technology is used in real work situations. These have been designed to enable our multi disciplinary Task Force to open up the so-called 'black box' of algorithmic decision-making and explore the legal[9] and social architecture that shapes their design and adoption.

The Equality Act 2010 and the Data Protection Act 2018 are the UK's main legal frameworks directly governing algorithmic accountability. But our research suggests that these legal frameworks – and how they work together – are not widely understood. So, the ETF, chaired by Helen Mountfield QC, has tested the application of these legal frameworks against our case studies. We have also used our case studies to discuss the nature of the problems that data-driven technologies present, the actions needed by key players to address them, and how to bring about these actions.

We have considered the underlying goals we wish to achieve,[10] and how law and regulation can contribute to achieving this and shape a better future of work. The high-stakes context of work, and the central role work plays in people's lives, means that focusing on fair use of AI at work should contribute to the broader debate on AI ethics and regulation too.

Our case studies

Please see our publication Machine Learning Case Studies for further details.

Thor

The Thor case study is a hypothetical example of use of AI to help hiring decisions. It 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.

Networkz

The Networkz case study looks at software which targets job advertisements at particular groups. It 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.

Brill

The Brill case study is an example of use of AI in performance management. It draws out pressing challenges arising from automated employee management, monitoring and enforcing productivity in order to save costs.

Introduction

Supported by Freshfields Employment Law team and IFOW's research network, we have also:

- Curated 4 ETF dialogues in which intersecting legal and institutional rules, roles and responsibilities to address the challenges presented in the case studies were evaluated
- Hosted an ETF workshop led by Dr Logan Graham on how machine learning works
- Co-hosted an open workshop on Equality and AI with the Institution of Engineering and Technology
- Undertaken new research on hiring systems, job advertisement and automated management at work
- Considered an evidence review undertaken and generously provided by the EHRC
- Considered member surveys undertaken and shared by ETF members Prospect and the CIPD
- Analysed a new technology survey undertaken by USDAW
- Undertaken and reviewed a public consultation on IFOW's protype equality impact assessment
- Undertaken a gap analysis of the Equality Acts 2006 and 2010.[11]

Several of our ETF members have also independently undertaken work which is relevant to our Terms of Reference (ToR), some of which is ongoing, which has informed our recommendations. We are very grateful to our Task Force members for generously sharing their work, ideas, and practical experience. We note, in particular, that Dr Reuben Binns has led and coordinated ICO's Guidance on auditing and AI, which has now been published; Professor Helen Margetts is working on updated guidance with co-author Dr David Leslie at the Turing Institute; and Joshua Simons is co-authoring a paper[12] on the introduction of positive duties to advance equality in the governance of machine learning.

Through our work this year, we have developed three key propositions, upon which this report rests:

First, data-driven technologies, and the data they produce, are the product of human agency. Human beings set the parameters for artificial intelligence, machine learning and other mass data systems. In particular, they design data-driven systems and they decide how to deploy the data that those systems generate.

Second, the products of data-driven technologies are not neutral; without careful scrutiny they have a tendency to reinforce and reproduce historic inequalities. This can adversely affect the aspirations and opportunities of millions of people.

Third, the existing framework of regulation, in this area, is inadequate to promote equality and fair play now, and into the future. The challenges we identify sit at the interface of data protection and equality law, and are not adequately met by either framework. Greater accountability in this area will require fresh legislation.

1

The myth of neutrality

Part 1

The myth of neutrality

Algorithms, artificial intelligence (AI), and machine learning (ML) are tools for using data to make or inform decisions. They leverage the patterns and regularities in data to make predictions, used to support or supplant human decisions.

Humans have used data to inform decisions for millennia. But as more of our behaviour is recorded in data, and more of our world is connected via the internet of things, the incentives to use of data in decision-making are becoming both stronger and more pervasive.[13] There is a remarkably widespread and persistent belief that unlike humans, who are 'inherently and inescapably biased',[14] these systems offer the possibility of a kind of neutrality and objectivity.

Reflecting this, a great deal of the literature promoting data-driven tools for use in the workplace hold out the promise that data-driven decision-making can avoid the prejudices and fallibilities of human cognition.

Our analysis of case studies made clear how flawed as well as how pervasive this myth is. There is nothing inevitable about how data-driven technologies will transform our society. It is the choices that humans make about how those technologies are designed and deployed in decision-making systems, not anything intrinsic to the technologies themselves, that shape who technology benefits and who it harms, which values it promotes and which it erodes.

In fact, our analysis suggests a point that is often under-appreciated by policy-makers: without intervention and oversight, the natural state of data-driven technologies is to replicate past patterns of structural inequality encoded in data, and to project them into the future.[15]

Data encodes the inequalities of the past

Data is any kind of information relating to a person, group, corporation, or other subject, most often but not always in numerical form. A dataset is always assembled for a particular purpose, and that purpose shapes how, when, and by whom data is measured, recorded, and collated. Data is useful because it captures common relationships among people, groups, corporations, or other subjects, including persistent patterns of outcomes. These relationships are useful because past patterns of behaviour can be used to understand and inform human decision-making.

But this means we must be careful to ensure that past injustices are not compounded and 'super-charged' by data-driven decision-making technologies.

Part 1
The myth of neutrality

The reason machine learning cannot be neutral is that data, on which machine learning models are trained, encodes history. The statistical relationships ML models learn to use almost always encode patterns of inequality and disadvantage, even when protected characteristics are excluded from datasets. The process of machine learning often uncovers the complex ways in which race, gender, class, and geography relate and condition the opportunities people are afforded. The data patterns in the data sets and sources, which inform the statistical systems, can ossify injustices and inequalities by presuming that outcomes are duplicated endlessly. They do not recognise that individuals and groups could perform better over time.[18] As a result, the data sets reflect the structures of human behaviour and power which produce them. Choices about how to use that data to make decisions cannot be neutral and humans must decide which of them are useful in identifying and predicting future aptitudes, and which demonstrate the unjustified impact of past patterns of privilege, which block fair recognition and harnessing of potential.

These choices increasingly shape access to work, the full range of employer functions, and new 'augmented' advisory and predictive functions.[19] For example, in the Thor case study an ML tool is used to rank candidates for a job from 1–10 based on their 'expressive skills', comparing CV, LinkedIn and social media data. Whilst the model excludes traits such as ethnicity, age, gender, it nonetheless exclusively recommends male candidates. This is because the model learned that verbs like 'execute' and 'dominate', which were used more often by male than female candidates, predicted future job performance. The reason was that most of Thor's existing engineers were men, and this language typified their CVs. The model simply learned to repeat the patterns of inequality that had previously characterised Thor's hiring system.

Understanding the technology

Algorithm

An algorithm describes any structured process for solving a defined problem. Algorithms have been used in complex decision-making processes throughout human history. Whilst algorithms are increasingly automated, they need not be. Algorithms include the rules welfare officials use to decide whether an unemployed person is eligible for universal credit, or the decision tree a bank clerk uses to determine who qualifies for a loan to start a new business, or the criteria to determine the distribution of grades at a particular school. Traditionally, each step in the decision process is explicitly stated and defined by humans.

Artificial intelligence

Artificial intelligence (AI), by contrast, is best understood not as a single technology, but as a scientific field although it can be used as a marketing term for a range of technologies. Machine learning (ML) is a category of AI in which computers learn from data how to accurately perform well-defined tasks, through 'experience.'[16] The 'experience' from which the computer program learns is almost always large volumes of numerical data.

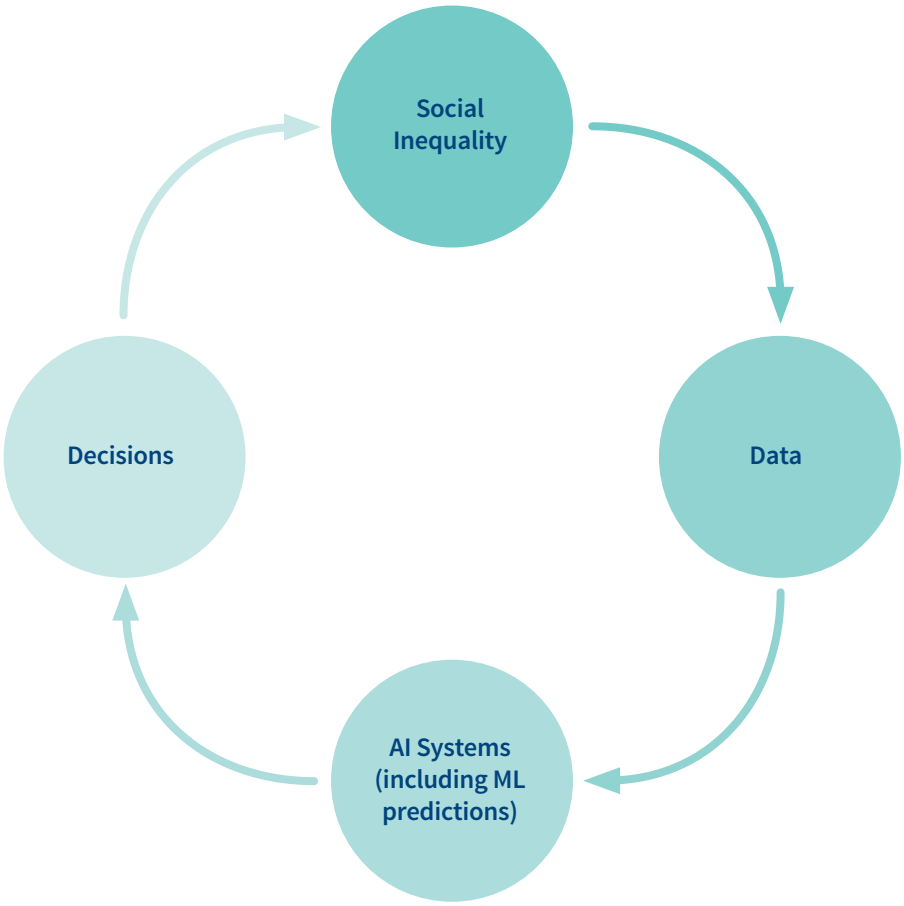
Machine learning

Machine learning is a statistical process in which data is used to train a model to make accurate predictions. ML systems learn from data which combination of statistically-related attributes most accurately predicts a particular outcome. Not every step in the decision process is explicitly stated by humans, but the process of machine learning involves a series of choices made by engineers and data scientists, managers and executives, embedded within particular organisations. Unlike traditional algorithms ML is an aggregation of different algorithms which are constantly redesigned in relation to each other, to achieve outcomes set by an initial algorithm shaped by humans.[17]

In the job advert recommendation system in the Networkz case study, the system shows men job adverts with higher average incomes than those shown to women. Because gender stereotypes are encoded in online behaviour, machine learning models reflect these patterns, and in using them to distribute adverts, serve to further entrench them by limiting the chances that women will see them. The larger the scale at which Networkz's job recommendation system operates, the greater the compounding effect the machine learning models that power it can have.[20]

This does not happen because machine learning has gone wrong, because a model is biased or a dataset is unrepresentative, but happens when it works exactly as designed.[21] The problem is not what the machine learning shows, but the assumption that this is necessarily 'right' or objective and should continue to happen in the future. This can create a powerful feedback loop, amplifying and entrenching social inequalities and systemic patterns of disadvantage. The quest for neutrality, in other words, reinforces the status quo. Resisting this demands proactive efforts.

The big feedback loop



Part 1
The myth of neutrality

The dimensions of data are increasing

In contrast to simpler algorithms (such as those used by Ofqual), ML systems identify a mind-boggling number and range of statistical relationships to predict their outcomes. These relationships need not make sense to the human mind, they need only support an 'accurate' prediction of the outcome a model is trained to predict. This means that ML can unearth dimensions of inequality that do not always fit neatly within existing understandings of the dimensions of inequality, but it can also make more focused discrimination within groups conventionally considered to be at a disadvantage.

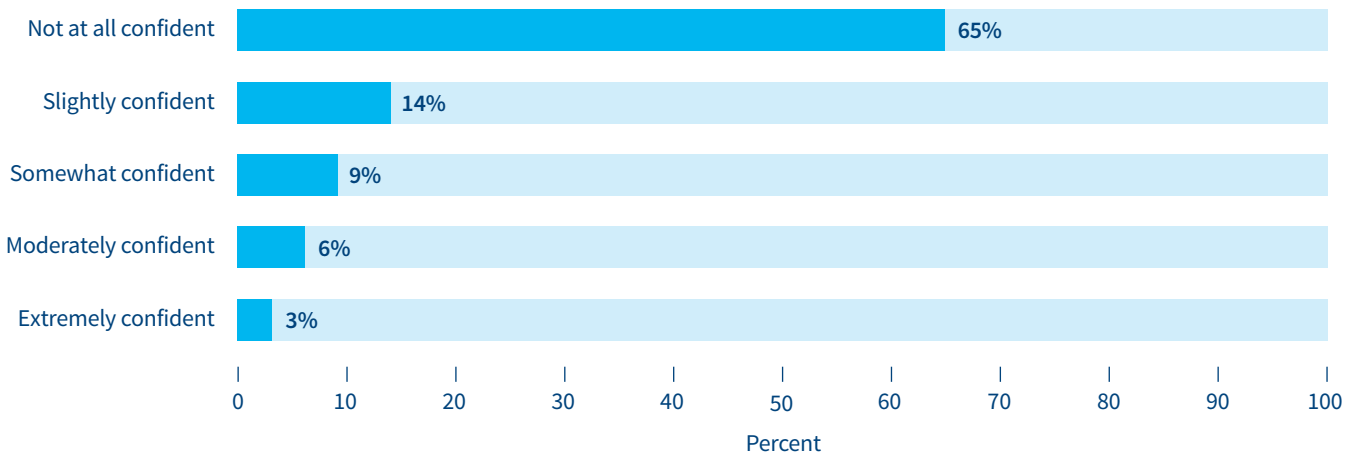
This problem of ML models drawing on lurking patterns of inequality to intensify discrimination is only made worse by the opacity of complex, high-dimensional algorithms. The more data variables that are included in an ML system's architecture, the less understandable the rationale behind its results will likely be. Likewise, in complex algorithms that draw patterns from non-linear connections between multiple variables, the relationship between inputs and outputs can quickly become difficult to understand.

This means that multi variable correlations, which are replicating patterns of inequality and causing discriminatory impacts can become buried in the opaqueness of the resulting algorithmic "black box." Once a complex ML model loses its interpretability, which in the context of machine learning refers to the ability to explain or present in terms which are understandable to a human, the origins of the discriminatory harm that its use may inflict become very difficult to access.

In an application of the ML tool used in the Thor case study, customer service ratings of existing staff were correlated with wider datasets about existing staff to create a model of a 'good employee'. Customer feedback ratings were on average higher for those from ABC1 economic backgrounds.
This may, on human analysis, reflect the fact that customers of Thor were generally from the same economic position and preferred to talk to people who sound like them.
But working out the reason why depends on human analysis and insight. The algorithm identifies correlations, not causation.
This is a critical and often under-appreciated distinction. As a consequence of this, an ML tool which incorporated voice recordings, or social media data, may be able to filter out lower-income candidates based on language

I know whether my data is being shared with 3rd parties

Total n = 963. Fieldwork completed between August and October, 2020. By USDAW in partnership with IFOW.



recognition from their video applications (now increasingly common) or data points which signify this economic status from their social media profiles (such as subscription to certain magazines, or use of certain language). As tools like this add more and more data variables, the points at which patterns of inequality and discrimination enter into the analytics can become less comprehensible and ultimately hidden from view.

In addition to more advanced assessment and grouping of particular communities (for instance, of place linked to certain accents or of socioeconomic status, as indicated by postcode, use of specific platforms, and so on) is the level of granular differentiation between individuals from within a particular group, allowing for more intersectional decision-making. Far from 'blind' human reviews in traditional public sector recruitment where details such as name may be taken off the list, in an ML process of recruitment distinctions could be drawn between a well-educated black woman from a lower socioeconomic background, and a well-educated black woman from a higher socioeconomic background and make decisions on this basis.

Shaping the future in the image of the past…

It is problematic to conceive of predictions as accurate, when they are used in decision-making systems which effectively determine the outcome they predict. The increasing use of prediction to make decisions is cementing historical inequalities, by projecting them into the future.[22]

The computational power of ML is what delivers a significant part of its economic gains; capable of engineering precision in complex logistics operations spanning continents within a single program.[23] As systems are refined to optimise efficiency, work has become part of this precision operation. A large part of the debate around the predictive capacity of ML in work has focused on hiring tools, as do two of our case studies. But 'algorithmic management'[24], most commonly associated with conventional gig sectors and platforms such as Deliveroo, Uber and Amazon, is becoming pervasive across conventional sectors too.[25] Such tools use prediction to make decisions which restrict, recommend, record, rate, replace and reward workers and incorporate data from an increasingly diverse range of sources.[26]

Instead of evaluating what employees can do, the proliferation of data encourages employers to try to develop a representation of who employees *are*, then use that to predict what they *might* do (or be able to do) in the future, based on the correlation between this data-based representation of an individual, and the performance of other individuals who share the same representational characteristics. The use of data to predict individual and collective identities and future actions encourages decision-making that is based on comparisons between similar individuals and groups.

…At greater speed, on a bigger scale

Decisions made about recruitment, hiring, promotion or terms of work by a human are taken about individuals, incorporating objective and subjective factors. In contrast, as we have seen above, recommendations made by an ML system about individuals are made on the basis of the relationship between data based representations of that individual, and features which correlate with data-based representations of 'ideal' candidates (in hiring) or workers (in performance management), as represented (imperfectly) by the data the system holds about both groups. Our case studies suggest that in practice, machine learning is often used alongside human decision-making. However, in practice, there may not be a discussion between parties engaged in the process, or conscious design decision about what an 'ideal' candidate is, with the ML system constructing this.

The central problem that Networkz's job advert recommendation tool poses is not just that individual women see fewer job adverts for higher-income jobs, it is the risk that in moving to platform-based advertisements, reaching millions more users relative to previous forms of media, huge swathes of the labour market which are already disadvantaged as a demographic group, may be foreclosed from seeing these opportunities. The larger the scale at which Networkz's job recommendation system operates, the greater the compounding effect the machine learning models that power it can have.[27]

As a consequence of human design choices

Human choices about the design and deployment of data-driven technologies shape their effects on society: who they benefit and harm, which values they embed and which they corrode. The regulation of these technologies must surface those choices, interrogate their stakes, and subject them to appropriate structures of oversight and accountability.

Two common words often misdescribe the problem that machine learning poses for regulation. The first is that AI is a "black box" which can be impossible for humans to understand or interrogate. There is a lack of clarity in the press and in popular writings on the character of the AI "black box". On the one hand, this refers to the proprietary protectionism of firms that are attempting to safeguard their intellectual property by not disclosing details about their software and computer code. This intentional lack of transparency is often cast as financially prudent and strategically necessary in competitive innovation environments. But, it is also used to set up unjustifiable roadblocks to sensible regulatory oversight and to evade reasonable expectations about public-facing assurance of fair practices and non-discrimination.

On the other hand, the AI "black box" refers to the aforementioned barriers to understanding that complex algorithms pose. This kind of opaqueness often leads to the setting up of a different sort of unjustified obstruction to regulatory intervention. In these instances, the limitations of human-scale cognition are treated with a sense of defeatism, and decision-making human agency is then ceded to the "smarter," more complex character of high-dimensional ML systems. This perspective, however, is misguided—especially as it applies to human-impacting applications that are processing social and demographic data.

Part 1
The myth of neutrality

An ML system that becomes opaque to its designers and users may conceal discriminatory inferences, making them inaccessible to impacted individuals and auditors. This is a fundamental problem as it relates to reasonable expectations of responsible social action. That is, when a ML system is used in social contexts where real people are affected, it should be able to meet their reasonable expectations about fair and equal treatment. However, not only will a "black box" application not be able to do this, but the widely-accepted purpose of using these kinds of statistics-based systems as empirical support for evidence-based reasoning will be violated. Regardless of how designers and users prioritise the optimisation of ML systems for predictive accuracy, when these systems impact individuals and communities, their outputs should be rationally justifiable to all impacted parties.

The attribution of "black box" status to AI systems can also be misleading because it obscures the critical role that humans play in designing machine learning models. When these models are used in practice across a range of sectors, including in the workplace, they are always as part of a socio-technical system which involves technical, moral, legal and political choices. Those choices are the object of regulation that seeks to structure accountability in the public interest. So if AI applications are simply treated as impossible to understand and interrogate, the fact that their development, construction and implementation arise from human decisions and are therefore subject to rational oversight and criticism will be obscured. The design and use of such applications raise problems which need to be addressed in their social and ethical implications and regulated accordingly.[28]

The second is that machine learning "automates" human decisions, such that the essential choice about how to use machine learning is whether to replace human decision-makers. This obscures the multiple ways in which machine learning models can be integrated into existing decision-making processes. Our case studies revealed that in practice, machine learning is often used alongside human decision-making, to produce reports about employee performance, or to provide an initial ranking of candidates for employment. The policy debate must become more sophisticated in distinguishing the different ways humans and machines can and do work together because each requires different ways of structuring accountability.

We have identified seven critical choices in the design and deployment of data-driven decison-support technologies. Forms of bias and inequality can skew decision-making at any of these choices, or a combination of them.[29] We briefly identify and illustrate these choices using the Networkz case study, to sharpen the points of choice for which technology regulation must hold organisations accountable.

There is no neutral way to make these choices. Insofar as datasets reflect the inequalities and injustices of our society, decisions about how those technologies are built and used will necessarily benefit some people over others and promote some values over others. If data-driven technologies are built to optimise efficiency, to be neutral and blind rather than to deliberately advance equality, their predictions will replicate underlying patterns of inequality, and the use of those predictions by humans in decision-making will entrench structural inequality on an unprecedented scale and with unprecedented speed.[37]

The development of policy responses must pay careful attention to the human and relational elements of socio-technical systems: we can only develop appropriate accountability mechanisms if we identify and understand human roles.

Seven critical choices

1.
Outcome: Setting the agenda and impact evaluation

From the first moment that decisions are made about allocating resources and dedicating labour power to exploring AI innovation projects, human choices are made about use contexts and the kinds of technologies and policies to pursue in the cases under consideration. Evaluation of the impacts on the individuals and communities affected by them should start here. In all our case studies, organisations begin to explore how the adoption of a statistical tool would affect people of different races and ethnicities, genders, and social groups. As we later recommend, these evaluations should be revisted throughout the design lifecycle, at the point of integration into a decision-making system and throughout its deployment, until the technology is retired. This dynamic evaluation should be systematised and integrated into a cohesive system for structured accountability.[30]

2.
Problem formulation and outcome definition

Choices made about how to delimit the problem space to be addressed by an AI project, as well as how to define the outcome that the model is trying to predict, are critical components of the justifiability of any innovation process, and they involve significant measures of human judgment. Deciding on what an ML model will learn *to do*—what outcome it will be trained to predict—involves determining what sort of target variable or measurable proxy indicates that successful prediction. In the case of Networkz's jobs tool, this is the probability that someone will click on a job advert and the probability they will apply to a job given they have seen the advert.

3.
Data collection and use

Choices made about where and who to collect data from, how to collect it and how to clean, wrangle, and curate it, are all important socio-technical factors towards which regulatory oversight should direct its attention. As we discuss below, choices about training data are often the most important determinant of a model's predictive power, practical utility, and the patterns of outcomes it produces.[31] The training data for Networkz's jobs tool includes reams of data about users' online behaviour, from the groups they like and the friends they have, to which job adverts they tend to click.[32]

4.
Features/variable selection and engineering

The fourth key choice is the shortlisting of the most important variables to be included in a model, from those in the long list selected in step 2: which variables should be used to predict, rank, or classify the specified outcome. While these are selected by a human in an algorithm, in the process of machine learning, a model learns which statistical combinations of features most accurately predict an outcome. A person can then adjust the weightings or decide to include or exclude certain features. In practice, step 2 and 3 are iterative. Selecting features (the input variables to be included in a model) is a key human choice in the ML lifecycle. The shortlisting of possible features from the data collected has many downstream consequences. Choosing which variables are to be used to predict, rank, or classify the specified outcome also involves subjective and potentially contestable decisions about how to define the variable categories themselves (such as how one might characterise racial or gender groups and how one might organise relevant sub-groups into chosen categories).

Seven critical choices *continued*

In our Networkz case study, engineers must decide whether its models to optimise job adverts should include gender, ethnicity or other characteristics or behaviours as relevant features, and determine trade-offs. For example, excluding some variables may make the model blind, in a certain sense, but it may also make it both less accurate (capable of efficiently predicting an outcome) and less fair.[33]

5. Implementation

One of the most important choices about the deployment of a model is how its predictions should fit into a broader decision-making system. Responsible deployment must address questions about how to appropriately train and prepare users to adopt new and practice-shifting technologies, and also to recognise their limitations. The predictions of a trained machine learning model can be used to support decisions made by humans or they can be used to supplant human decisions. Networkz's jobs tool uses machine learning models to automatically determine which users see which adverts, but this does not have to be the case.[34]

6. Communication of model predictions and limitations

When a model is used to support rather than replace human decisions, the way in which a model's outputs are presented and communicated affects how humans use it.[35] Models might display a numerical risk score with colours, green for lower risk levels and red for higher risk levels. The threshold at which risk scores are presented as green, yellow and red may significantly affect how people use those predictions to make decisions. This must be understood and implementers must be trained to identify the limitations of the model and statistical generalisation more generally (including error rates and uncertainty).

7. Making runtime adjustments

If problematic disparities are detected during runtime reassessment of algorithmic impacts, organisations can then deploy a range of mitigating measures that we have considered elsewhere.[36] It is telling that auditing and 'adjustment' tools do not adequately consider impacts on equality and are, generally, not designed or equipped to address many forms of bias, discrimination and inequality when they are detected. Imposing particular statistical definitions of fairness, in which equality can be seen as niche interpretation, is generally not the most appropriate way to make adjustments to counter or mitigate adverse impacts on equality. While mathematical definitions of fairness can be useful, focusing on them often obscures both the material preconditions of equity and the prior human choices about how machine learning systems are designed and adopted, which we think have been given inadequate attention by policy-makers.

2

Structuring accountability

Part 2

Structuring accountability

Accountability is central to democracy, and structuring accountability of private powers to the public good is a critical function of the legislature.

Precisely because there is no neutral way to design and deploy data-driven decision-making tools, it follows that we, as a society, must decide the goals and values that should be built into their design and how these tools should and should not be used. This invites consideration of the ways in which these systems can cause or compound social harms, so that we can clearly articulate those harms which are unacceptable and we wish to prevent.

This requires forms of accountability to be structured in the public interest, ensuring that people embedded within organisations, and the organisations themselves, are held to account for choices about the design and deployment of technology that shape work and society. This is needed to affirm human agency at both the individual and collective level.

However, we have identified a series of institutional obstacles to achieving this goal. We discuss them in this section, before we move on to an analysis of our legal ecosystem and the proposal of a path forward.

Decisions are diffuse and span multiple organisations

Statistical decision-support systems, and machine learning in particular, make power both more diffuse and more concentrated.

Machine learning makes power more diffuse because the choices which shape the effects of decision-making systems are distributed across a wider range of organisations. The range of organisations in which the human choices about the design and deployment of data-driven technologies are made often cut across the traditional boundaries of regulation: public and private, large and small, and different sectors like AI development, retail, and internet platforms. Drawing up a workable regulatory framework means making choices about where responsibility will be imposed and how accountability will be enforced.

For example, Thor's hiring system involved a combination of statistical tools developed in-house, those procured from other companies, and data purchased from data brokers. Given that the patterns of inequality produced by Thor's system result from a combination of all these components, who should be responsible for the disparate impact of Thor's system across gender and socioeconomic class? Or should each decision-maker play a part?

Part 2
Structuring accountability

We must also bear in mind that because machine learning increases the scale and speed at which decisions can be made, machine learning also concentrates power. The design choices of a few people can shape the lives of more people more quickly than ever before, even though those people are often distributed across several organisations in multiple sectors. This presents a particular challenge, because the computer scientists and software engineers who design machine learning models do not always realise the social, moral and political stakes of the choices they make.[38] Unpicking this challenge is further complicated by potential intellectual property rights, which we return to below.

The challenges that statistical tools present to accountability, and to safeguarding equality, are rooted in choices made by people. What makes the challenge of structuring accountability hard is that these choices involve the exercise of a concentrated form of power that is distributed across multiple organisations in several sectors.[39]

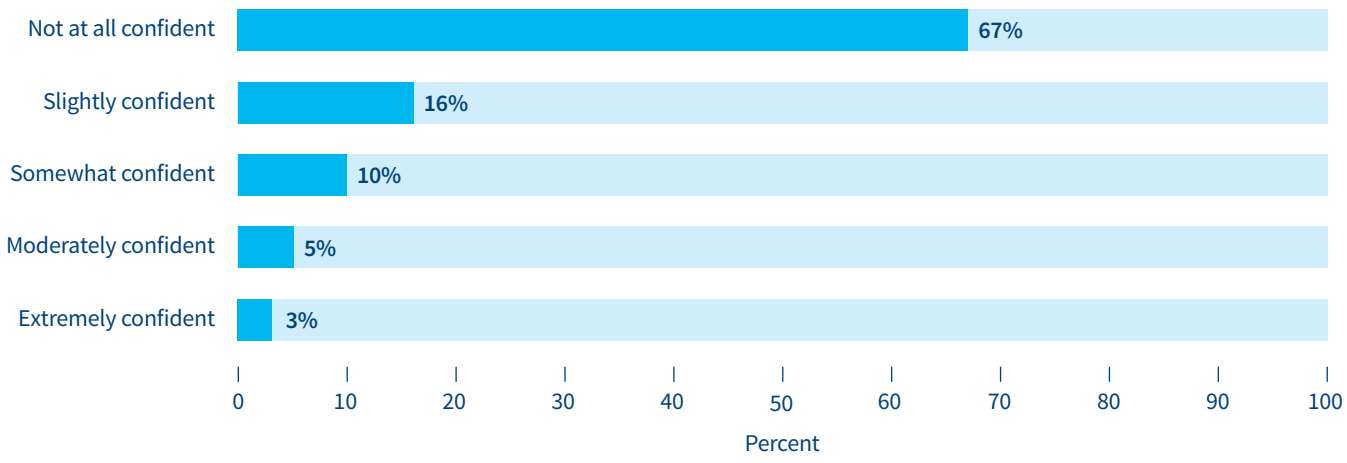
The language of statistics obscures human roles

Similarly, those who may be broadly accountable for the effects of these choices, whether corporate managers or CEOs, public officials or managers, often do not understand the language of computer science in which these choices are articulated, or what interests and values are at stake. This can obscure human roles and obstruct public and private conversations about meaningful forms of responsibility.

In particular, the stakes of human choices that shape the effects of data-driven technologies are often hard to identify, let alone interrogate, because those choices are often articulated in the language of statistics.[40]

When statistics are invoked with an air of certainty and scientific neutrality, without the uncertainty and humility they deserve, they can bury choices that implicate fundamental values about which reasonable people disagree, and that prioritise the interests of different groups in society.

If my data is used to assess or make predictions about my performance, I know how it is used to do so
Total n = 974. Fieldwork completed between August and October, 2020. By USDAW in partnership with IFOW.



Part 2
Structuring accountability

In the Networkz case study, when click probability is unevenly distributed across men and women for high and low-income job adverts, the choice to predict click probability *sounds* like a technical choice but is, as we later demonstrate, exactly the kind of choice that must be interrogated.

This not only makes it harder to apply existing laws to data-driven technologies, and to identify gaps, but it can also stymie public debate about what values and interests should be prioritised. This can make it hard to articulate the nature and extent of the accountability challenges we face or build consensus about what to do.

This problem can be seen even within HR departments, like those in the Networkz and Thor case studies. While addressing these kinds of questions could effectively be part of their reinvention, professional development has not yet kept pace with the rapidly-accelerating challenge they are presented with.[41] In sum, not only do these choices cut across organisations, but different organisations and different departments within organisations hold different capabilities to understand and evaluate human design choices.

Information asymmetry is growing

Data-driven technologies tend to exacerbate existing asymmetries of information. Information asymmetry means a failure or imbalance of accessible information between people, groups and organisations. Technologies driven by increasingly complex data sets simultaneously demand highly personal information and share less about their own analysis and use of it. It follows that there is an increasingly uneven playing field between those who have an interest in shaping the design and deployment of statistical tools, and those experiencing the sharp end of their use.

There are few clearer examples than the asymmetry of information between worker and employer. Not only do many workers lack the familiarity with statistics or computer science required to understand and evaluate the statistical tools used to make decisions about them, they often have little knowledge that statistical tools are even being used at all or how they are designed.[42] Because such information is rarely publicly shared by either the private or public sector, and when it is, rarely explained with the necessary clarity and simplicity, workers' lives are being shaped by statistical decision-making process about which they have extremely limited information or control.[43] Further, data-driven systems are fed by increasingly complex, pervasive and invasive data sets that reach outside the traditional workspace.[44]

In these circumstances, it is not surprising that the new USDAW technology survey found that only 52% were 'not at all confident' that they knew why and for what purposes information was collected about them; and that 67% were 'not at all confident' that they knew their data was being used to access or make predictions about their performance.[45] This lack of transparency about the purposes for which data is being collected and processed, compounded by the application of ML technologies, constrains not only individual but public understanding and, in turn, policy response.

If people are unaware of what information is being used to judge their performance, or what inferences are being drawn to determine their access to and terms and conditions of work, they cannot begin to determine or challenge whether their treatment is fair or otherwise. This is further exacerbated where workforces are remote or disaggregated, for instance when working for a platform, or working away from a central office, which we anticipate will become increasingly common.[46]

Part 2
Structuring accountability

There are other areas in which information asymmetries are being exacerbated as well. For instance, those responsible for the effects of data-driven technologies, including on workers, often have limited understanding of the technologies or design choices for which they are ultimately responsible,[47] whether unions, business managers and CEOs, civil servants or ministers responsible for public bodies.

Regulators without existing expertise in interrogating data-driven technologies often lack both the information, and capacity to analyse that information, to evaluate whether those technologies respect existing laws and regulations. For instance, while the EHRC have deep expertise in the content and enforcement of UK equality law, they will need to collaborate with other regulators like the ICO to apply that understanding to data-driven technologies. We return this in Part 4.

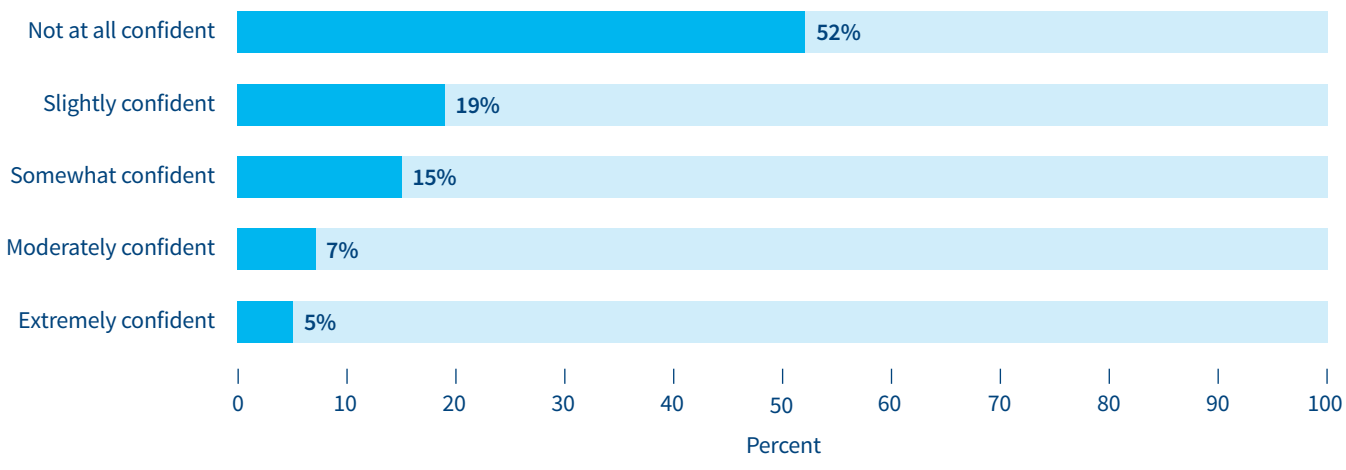
In some complex forms of machine learning and artificial intelligence, even computer and data scientists and engineers may not wholly understand *why* a particular system generates the predictions it does, although they may be equipped to analyse a system to answer the most important questions about it for good governance and legal compliance.[48]

Perhaps most fundamentally of all, asymmetries of information are starkest from the perspective of citizens. Workers and citizens in a democracy have a strong interest in living in a world they understand, whose decision-making processes they can interrogate and contest, and whose fundamental principles they can understand and debate with others. Without appropriate structures of oversight and accountability, data-driven technologies can undermine this fundamental goal.

Each of these forms of asymmetry of information must be addressed in different ways. The kind of information disclosure required to empower citizens to engage with the decision-making systems which shape their lives will be different to the kind of information disclosure required to enable regulators to effectively monitor and enforce obligations. While there may be no single silver bullet to address these asymmetries of information, it is clear that the purpose, motivations and rationale of organisations matter just as much as the statistical logic of how a system generates predictions.[49]

I know why and for what purposes my employer uses data collected about me

Total n = 977. Fieldwork completed between August and October, 2020. By USDAW in partnership with IFOW.



Similarly, examining the reason *why* information is asymmetric, and whether a particular asymmetry could have been mitigated or avoided, is relevant to achieving higher levels of transparency.[50] In Part 4 we look at how the significant steps made in technical explainability of algorithms provide a sound basis to extend existing transparency obligations to the 'socio' aspects of these socio-technological systems.[51]

Privacy is an insufficient frame for accountability

Interpretations of data justice often respond to limited perspectives on the societal risks of data-driven technologies, with efficiency and security on the one hand and concerns about privacy and data protection on the other.[52] These approaches and frameworks for accountability each make some inroads but do not squarely address the core challenges we have identified, which fall at the interface between our regimes for data protection and equality. We return to this below.

In particular, the GDPR (incorporated into UK law under the Data Protection Act 2018) is widely-considered[53] to have been designed in part as a tool for structuring accountability over the design and deployment of algorithms. While data protection aims to protect fundamental rights beyond privacy, including non-discrimination, it primarily addresses risks to those rights as they arise from the processing of personal data.[54] As a result, its regulatory tools are generally limited by its focus on conditions around the processing of personal data rather than a holistic assessment of the whole socio-technological system.

Nonetheless, the GDPR does contain two 'proactive' tools in particular, which address aspects of the design and deployment of AI systems, alongside *ex ante* obligations[55] to provide information about basis and logic involved in processing personal data.[56]

Data Protection Impact Assessments are required by Article 35 of the GDPR where processing of personal data is likely to create high risks to fundamental rights and freedoms of data subjects (including privacy, but also equality). Where those high risks cannot be sufficiently mitigated by the data controller, the DPIA must be referred to the ICO for prior consultation before proceeding. Second, individuals may make an express claim under Article 15 of the GDPR to receive a copy of their personal data, for instance where they suspect foul play. We return to these in Part 3.[57] These are both 'proactive' in the sense that they can be triggered by the actions of a data controller or data subject rather than as part of a reaction by the regulator.

AI ethics are a stepping stone but not a substitute for technology regulation

There is widespread agreement that "AI Ethics" and the development of now over 80 proposed AI ethics frameworks[58] is not an adequate approach to regulating and governing data-driven decision-making systems. Whilst it is important to avoid the conflating the way private companies and governments can use the moral rhetoric of ethical codes of conduct, and non-binding, high-level principles as an evasive means of ethics washing, we note the constructive role that ethical values can play as providing a normative basis for the codification of principles-based standards and regulation.

In particular, dialogue around AI ethics has resulted in a growing consensus about what kinds of ethical concepts must be included in any technology governance approach.[59] This has triggered a new momentum among technical standards-setting agencies to widen their remit and take an ethical and socio-technical turn.[60] For example, the Institute of Electrical and Electronics Engineers (IEEE) has recently formed the P7000 series of standards and certifications (now in production),[61]

Part 2
Structuring accountability

building on their Ethically Aligned Design guidance, and the International Standards Organisation (ISO)[62] has recently formed[63] a working group on the Trustworthiness of Artificial Intelligence, which has begun to think through how to construct standards for "characteristics of trustworthiness, such as accountability, bias, controllability, explainability, privacy, robustness, resilience, safety and security." The OECD AI principles, to which the UK is a signatory also identify five complimentary values-based principles for responsible stewardship of AI.

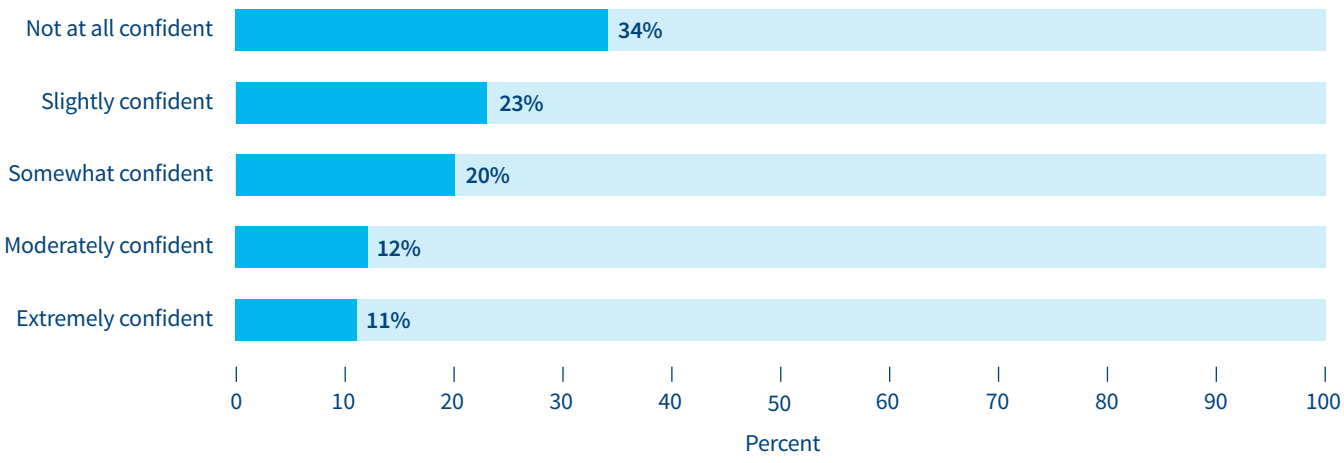
In the UK, the ICO has teamed up with the Alan Turing Institute to produce a guidance on AI explainability, *Explaining decisions made with AI* (2020).[64] Drawing on the ethical values and practical principles laid out in the UK's official public sector guidance on safe and ethical AI[65] the ICO/Turing approach to explainability calls on algorithm producers and users to address issues of ethical evaluation, responsibility, fairness and discrimination, over and above the narrow view that explaining the outcomes of AI decision-making systems involves the presentation of their technical logic. The Centre for Data Ethics and Innovation is also working a Bias Review due to be published later this year.

These indicators of the movement from principles to practice signal a shift towards increasing codification and regulation, building on this conceptual groundwork. Taken together, ethical guidance, practice-based principles and professional standards create normative vocabularies upon which new legal frameworks for accountability can draw.

Notwithstanding this progress, our work has identified a clear gap, in terms of recognition and attention: equality impacts. Whilst we recognise the growing body of work[66] that is beginning to integrate ethical concepts into standards and regulatory thinking, our case studies and workshops have emphasised the need for clearer, and harder, direction in how to safeguard *equality* in the design and use of data-driven tools. There is widespread misunderstanding of the risks that data-driven systems pose for equality, and a lack of confidence among business owners as to how they can act responsibly to avoid them, especially in the workplace.[67] Non-binding,[68] high-level principles have not been able to provide this direction, and are not changing design choices in practice.[69]

I trust my employer knows how to protect my rights when using my data

Total n = 940. Fieldwork completed between August and October, 2020. By USDAW in partnership with IFOW.



Part 2
Structuring accountability

In particular, there is limited understanding that there is no neutral or fail-safe way to build data-driven tools, because accurate and unbiased statistical models will reproduce existing patterns of inequality encoded in data. If we are to ensure these tools do not compound and 'supercharge' these inequalities, they must take this into account, make appropriate adjustments, and take steps to advance equality between the people and groups subject to their decision-making. This means tools must be "designed for equality" across the innovation cycle.

In spite of a lot of noise, this goal has not been met by corporate social responsibility. There are also common limitations and misapplications of market-based solutions, such as technical auditing tools, that we have explored elsewhere.[70] These limit, rather than encourage dialogue about collective adverse impacts, or what others are calling 'collective harms',[71] more broadly.[72] At the same time, many employers want to identify and nurture talent from the broadest possible pool, in a fair way, and we have observed an increasing wish by employers to understand the immediate and wider implications of AI and ML on equality and for clear direction on responsibilities throughout the innovation cycle to own, evaluate and address adverse impacts.[73]

The Networkz case study presents a job advertisement tool which was most efficient in getting clicks on adverts when it promoted lower-paid and conventionally 'female' jobs, such as those in care, to women, and higher-paid 'male jobs' such as engineering positions to men. Networkz decided they had a responsibility not to exacerbate existing patterns in user behaviour, but equally that they did not have a responsibility to alter those patterns. Their response was to adapt the model to determine a maximum 5% income gap between jobs advertised to men and women. But the result was reduced efficiency and reduced satisfaction across the board: the system showed people jobs which they were less likely to click on, apply for, and get if they did apply. In this context, engineers require help and guidance. Further, it should not be left for Networkz alone to determine their responsibility, or that there was only one method to mitigate what they knew was happening.

In order to ensure job advertising technology does not create undesirable new collective harms or inequalities of opportunity, technology regulation must structure accountability for the use of such software. This requires an exploration of who holds responsibility for responding to the patterns identified and reproduced by the system; and rules to clearly identify who is responsible for making those choices.

We note that some countries have begun to develop regulation to achieve this, such as the Algorithmic Accountability Act tabled in the U.S. Senate, the Data Protection Act being developed and implemented in India, as well as draft legislation in France, Germany, as well as the European Commission's work-in-progress Digital Services Act.[74] In October 2020, the U.S. House Judiciary Committee published an extensive report that argues for refocusing competition on market power and institutional *accountability* rather than consumer welfare, exploring how the major technology companies could fit within a comprehensive regulatory framework.[75]

By drawing on the best of these existing efforts, and avoiding some of the emerging pitfalls, the UK is well-placed to develop world-leading regulation to structure transparency and accountability in the governance of data-driven technologies.

3

The regulatory ecosystem

Part 3

The regulatory ecosystem

This section of the report examines the different legal regimes that already exist and might apply to the case studies set out above. This is a complex task. There are a variety of legal and institutional frameworks that are relevant to data-driven technologies.

Our core conclusion, developed below, is that existing statutory and regulatory frameworks do not provide sufficient accountability for the key individuals at the critical stage in the design and use of data-driven technologies.

In preparing this section of the report, we have engaged in a dialogue with technologists, regulators, lawyers, unions and business representatives. The purpose of this discussion was to identify and to hone in on the gaps within and between the various existing and overlapping regimes. We conclude that there are unnecessary complexities and other institutional obstacles to achieving meaningful accountability for the use of data-driven technology. The case studies – which we develop further below – demonstrate these flaws in practice. Any future regulatory solution must engage with the actual practices of businesses, government, regulators and civil society. Through our case studies, analysis has focused not only on the law, in abstract terms, but on the ecology of institutional actors, norms and practices through which the law is realised.

Use of algorithms, machine learning and AI engage a number of legal issues and legal frameworks. The primary legal spheres involved are data protection law, the Data Protection Act 2018 (which incorporates the General Data Protection Regulation (GDPR), as well as data protection regimes for law enforcement and intelligence services) as enforced by the Information Commissioner's Office (ICO); and equality law mostly contained in the Equality Act 2010 ("EA"), enforced by individuals or the Equality and Human Rights Commission (EHRC).[76] The Task Force also touched upon how and to what extent fair use of AI and ML at work were influenced by trade union law, as it relates to collective bargaining[77] and competition law.[78]

We provide, in Annex 1, a high-level overview of the key institutions in the UK that play a role in way that AI is used and regulated in the workplace, ranging from regulators to non-profit organisations. This highlights that there is no single regulatory or other body charged with overseeing the use of AI, ML and data-driven technologies in the workplace.

Part 3
The regulatory ecosystem

The Data Protection Act and GPDR

Data protection law is often seen as the primary regime for the oversight and regulation of technologies by employers, organisations and government. Data protection law is aimed at protecting people from *risks* arising *as a result* of the processing of their personal data, and offers *individuals* legal rights in prescribed circumstances. It is not aimed at countering novel forms of collective harm or the projection of group-based structural inequalities into the future.[79] Data protection law covers inferential analytics about protected groups insofar as they are based on analysis of personal data from individuals in those groups; and if they are applied to an individual to produce an inference about that individual, that inference is personal data.[80] While AI has typically been thought to be governed by intellectual property rights, data protection will also apply to the processing personal data involved in training and deployment, and will apply to AI models themselves where they contain personal data.[81] But analysis of our case studies shows that data protection law does not always offer sufficient focus on the harms to equality which can arise as a result of use of algorithms at work, nor adequate remedies to those adversely affected. Collective rights, access to relevant information or enforcement barely exist in the GPDR outside the requirement that representatives of data subjects are consulted on DPIAs.[82]

The UK's DP regime is a principles-led framework, governed by data protection principles which are intended to govern data processing:[83]

- Lawfulness, fairness and transparency
- Purpose limitation
- Accountability
- Data minimisation
- Accuracy
- Storage limitation
- Integrity and confidentiality (security)
- Accountability

Below we present aspects of data protection law which are relevant to the ETF ToR and case studies.

GPDR's role in design choices

As we have seen in Part 1, design choices shape the way ML works and the effects it has. GDPR sees a 'data controller' as any person, company, or other body that determines the purpose and means of personal data processing (this can be determined alone, or jointly with another person/company/body); or a 'data processor' if they do so under instruction of another data controller. The GPDR does not draw a clear distinction between the collection and the use of data.[84]

An important guiding principle in data protection law is *'data protection by design and by default.'* The ICO guidance on this principle states that data controllers and processors are required to put in place appropriate technical and organisational measures to implement data protection principles effectively, and safeguard individual rights. This means data protection principles must be 'baked in' to processing activities and business practices from the design stage, right through the lifecycle. Previously known as 'privacy by design', this was elevated from mere best practice to a mandatory part of data protection law under the GDPR. As we discuss below, there is no equivalent in the EA.

The GDPR generally prohibits data controllers from making decisions:

(a) that are based *solely* on automated processing; and

(b) that produce *legal effects* concerning an individual or similarly significantly affects an individual.[85]

The ICO suggests[86] that a process will not be considered 'solely' automated if someone weighs up and interprets the result of an automated decision before applying it to the individual: human involvement has to be 'active' and cannot be a token gesture.[87] But the extent to which an appropriate human review must be made before the decision, or has discretion to alter, is not stated.

The European Data Protection Board has confirmed that decisions that "similarly significantly affect" individuals include any decisions that have the potential to significantly affect the circumstances of the individual for a prolonged or permanent period, including impacting them financially.[88] This means that all the key decisions about recruitment, pay, terms or promotions featuring in our case studies are covered.

If an employer wishes to use a solely automated decision-making system that has legal or similarly significant effects, then the GDPR only allows this in the conditions:

(a) it is *necessary* for entering into or performing a contract between the individual and the controller (i.e. the contract of employment between the employee and the employer);

(b) it is *authorised* by law (e.g. fraud, tax-evasion monitoring); or

(c) it is based on the individual's *explicit consent*.[89]

Consent must be 'freely given' and the Article 29 Working Party takes the view (endorsed by the UK Information Commissioner) that it is unlikely that employees will be able to give their consent freely due to the inherent imbalance of power between employer and employee. In the Thor case study, current employees were asked to provide links to their social media profiles as part of their annual employee satisfaction survey. As a result, Thor has a database of information on public social media accounts, such as Facebook, Twitter and Instagram. But requesting consent did not come with a full explanation of the ends such an analysis could serve, and so would not provide a lawful basis to predict job tenure prediction (even if there had not been an imbalance of power).

In practice, employers are using the 'consent' ground less[90] and opting for the '*necessity for entering into/performing a contract*' to make decisions that fall under Article 22. In addition to imposing a restricted range of lawful bases for processing, Article 22 also includes a number of other safeguards, including the right of the data subject to obtain human intervention, to express their point of view and to contest the decision.

Further, such decisions cannot be based on 'special category data' under Article 9, unless the data subject has consented or processing is necessary for reasons of substantial public interest. Special category data includes data which reveals a range of categories which significantly overlap with characteristics protected under the Equality Act. Neither of these grounds are likely to apply in our case studies because consent cannot be 'freely given' where there is an imbalance of power between the controller and data subject, and ordinarily data processing by an employer would not be to advance a 'substantial' public interest.

Often, organisations like Thor may seek to avoid the more stringent requirements of Article 22 altogether by claiming that the decisions they make using algorithmic systems are not 'solely automated', but rather that the algorithm's output is just one factor weighed up by a human decision-maker. In such cases, a less restricted range of lawful bases may be available under Article 6(1), including that it is necessary for the purposes the 'legitimate interests' of the data controller. This ground requires the controller to establish that the processing is *necessary* for their own purposes or those of a third party, and to perform a balancing test to assess whether those interests are overridden by the interests or fundamental rights and freedoms of the data subject. This is relevant to our exploration of the similar proportionality test under the Equality Act, below.

In so far as it creates restrictions around the use of 'solely' automated decisions, the GDPR underpins growing arguments for the development of a 'human in command' and consultative approach, which was initially advocated by the European Economic and Social Committee and has been endorsed by the ILO and UNI Global Union.[91]

Part 3
The regulatory ecosystem

Using Data Protection Law to understand how a decision was reached

Without transparency regarding the purpose of a system, approach taken and extent of automated-decision making, it is very difficult for workers or their representatives to understand whether decisions taken are fair, or to apply other legal regimes.

Some limits of the 'transparency' provisions are yet to be tested in the context of work in a court of law. However, pre-empting this, the Task Force has tested the two main regulatory levers enabling sharing of information about ML design choices in the GDPR against our case studies. These are proactive, one ex ante and the other ex post facto.

First, Article 35 'enhances data controller responsibility' by requiring businesses to undertake a data protection impact assessment (DPIA/PIA) where the processing activities they are conducting are likely to be 'high risk' to the rights and freedoms of individuals.[92] DPIAs emerged in the 1980s, initially as a purely voluntary measure, and have since become a requirement under GDPR, where businesses determine the processing of sensitive data to be 'high risk'. The 'high risk' threshold covers each key decision in our case studies, although this is not widely understood.[93]

When a DPIA reveals that a project is high risk, even after risk mitigation measures have been identified, data controllers must consult with the relevant national supervisory authority (in the UK context, the ICO). This approach is a hybrid model of self-regulation, command-and-control regulation, and co-regulation. A form of combined or 'meta'-regulation, the State tries to make corporations responsible for their own regulation to save resources.[94] A framework and methodology for the assessment process is not specified by the statute.

On the upside, DPIAs are an important tool for evaluating the impacts on the rights and freedoms of data subjects of decisions 'ex-ante': they must be completed before processing starts and then continually assessed. This could shape approaches to optimisation criteria or invite reflection on questions such as whether is it *fair* to ask for friendliness in an engineer. But DPIAs are ultimately procedural mechanisms which focus on measures to reduce the risks to individuals' rights and freedoms resulting from the processing of their data. While DPIAs are expected to address the risk that 'processing may give rise to discrimination', the relative absence of any focus on equality in DPIAs, for private as well as public bodies, limits their potential as a regulatory safeguard from the specific problems we have identified in this report.

Even if a controller identifies and mitigates risks of discrimination against data subjects, such mitigation measures are not expected to entirely eliminate such risks. And there may not be a direct and obvious relationship between the way the *personal data is processed* and the consequences for discrimination. As a result:

'Cases will fall between the cracks of each framework – if you do a DPIA you have to mitigate risks to fundamental rights and freedoms, one of them being non-discrimination, but the penalty for not doing a DPIA is not going to be the same as breaching equalities law – which is right given it's the difference between conducting a procedure, and breaching a fundamental right.'

Equality Task Force member

Most importantly for our purposes, these assessments are not required to be made public, even to people affected by them, and so cannot be scrutinised by bodies who wish to understand the impacts of these decision-making systems in practice. We note, however, that some responsible employers are choosing to disclose these assessments on a voluntary basis.[95]

'While it makes sense to use DPIAs to advance equality considerations, we need to think about this in relation to what employers do and don't know about the potential impact for discrimination.'
Equality Task Force member

A second mechanism making aspects of processing transparent is the Right of Access (Article 15 (3)) GDPR. In the DPA this right is enacted through individual 'Subject Access Requests', which are submitted to controllers (e.g. current or prospective employers) by individuals who are typically either curious, or suspect foul play after processing has started. The right of access affords the data subject the opportunity to obtain information from the controller about any processing of personal data concerning them; essentially this is the same information that the controller is already obliged to provide under Articles 13 and 14, which includes access to meaningful information about the '*logic of processing,*' if Article 22 applies.[96] In addition, Article 15 provides the right to obtain a copy of their personal data. This would include a *copy* of any personal data generated by an algorithm, such as predictions about their job tenure in the Thor case study, or productivity evaluations generated by the Brill system. This should provide contextual information about the operation of an algorithm, with respect to an individual.

But Subject Access Requests have several limitations too. They do not have a standard format and are tasked only with revealing information relevant to an individual, so they cannot reveal patterns of discrimination, affecting group level outcomes. In particular, group and/or union rights to access this information, which would be required to understand collective and relative impacts, are not established; and the socio- and human aspects of the decision-making process are excluded, because the DPA focuses on the mechanics of the technology and perspective of the individual.[97] And, as Professor Lilian Edwards has pointed out, the right to meaningful information about the logic of processing, once it is triggered, is restricted in type and dimension.[98]

'Employers' aren't aware of the information they need, there's no duty to gather it, and in turn trade unions can't see it'
Equality Task Force member

Enforcement is another problem. A recent study of 150 apps and 120 websites found that less than half of access requests were answered satisfactorily, and around 20% disclosed personal data to impostors.[99] As these particular requests refer to data protection rights which are now over two years old, this suggests the regulatory environment, even for individual requests, is not able to ensure fair and transparent process and outcome, without additional support.[100] We return to enforcement by the ICO below.

We note the general prohibition on collection of special category data under Article 9, which overlap with characteristics protected under the Equality Act. This is often misunderstood or used as a shield against keeping data on protected characteristics, which is generally required to monitor and evaluate equality impacts, boosting the case for explicit obligations and rights imposed by law.[101] But Article 9 allows processing of special category data where necessary for reasons of substantial public interest; and Schedule 1(8) DPA specifically permits the processing of special category data with a view to enabling equality to be promoted or maintained.

The Equality Act 2010

The Equality Act 2010 ("EA") is the consolidating statute which contains the UK framework of equality rights and obligations. Its object is to prevent unlawful discrimination and to secure 'better outcomes for those who experience disadvantage.'[102] The Equality and Human Rights Commission describes its focus as being to 'protect the *rights of individuals* and advance *equality of opportunity* for all.'

The EA uses a variety of mechanisms to apply the ancient principle of equal treatment that 'like cases are treated alike and unlike cases are treated differently', and that unlike cases should be treated 'in proportion to their unlikeness'.[103]

The first two provisions are negative, prohibiting organisations and individuals from certain forms of conduct, whilst the second two are positive, requiring the organisations to whom they apply to address specific equality-related goals. We focus on the first two provisions here and return to the second in Part 4.

In the fields where it applies, the EA:

Prohibits direct discrimination 'because of' a protected characteristic (like race or sex) unless a specific statutory defence is given.

For example:
A police force could not discriminate for or against women in employment, unless there was a genuine occupational qualification (perhaps in first response to victims of domestic violence).

Prohibits indirect discrimination – a provision, practice or criterion which has a particular adverse effect on a particular group and which cannot be proportionately objectively justified.

For example:
A police force could not lawfully use a keen interest in playing football as a criterion for a job.

Requires decision makers to make reasonable adjustments for disability.

For example:
An employer may have to adjust the on-the-job training for trainees with physical impairments and permit them to undertake roles which do not require the same degree of physical mobility as other trainees.

Requires public bodies to give due regard to the need to avoid unlawful discrimination, advance equality of opportunity and promote good relationships between members of other groups.

For example:
A local authority would be required to consider the equality impact of changes to the means by which council housing is allocated, including the use of an algorithm to identify who might be eligible and prioritise among cases.

Direct discrimination

Direct discrimination is defined in Section 13 of the EA: "A person (A) discriminates against another (B) if, *because of a protected characteristic,* A treats B less favourably than A treats or would treat others." Direct discrimination concerns unfavourable or less favourable treatment of individuals in the same or similar situations, by either an act or an omission to act, '*because of*' a protected characteristic. So, the rule against direct discrimination aims to achieve *formal equality of treatment*: there must be no less favourable treatment between otherwise similarly situated people because of protected characteristics.

Underlying motive is irrelevant to direct discrimination,[104] as it is only necessary to enquire *why* a complainant has experienced less favourable treatment.[105] 'Because of' is an objective causal test so it does not matter if a person's motives were non-discriminatory or even laudable – if the less favourable treatment is *because of* the protected characteristic, it will constitute direct discrimination. This is important because choices about the design and deployment of automated and semi-automated systems are often buried in the technical language of computer science.[106]

'From an equality law point of view, if an employer can't explain why they recruited someone, that's equivalent to saying 'they're just not my kind of person'
Equality Task Force member

Data-driven technologies can directly discriminate against the individuals whose data they process. An algorithm might be designed to 'treat' women, the disabled or children less favourably because they are less 'desirable' consumers, for example. Coders, designers and programmers must be aware that the decisions they take about what value to place on certain characteristics can have the direct effect of treating some groups less favourably. Decision-makers – themselves blind to the design of the data-processing system – may then rely on the outputs as justifying their own unequal treatment:

'When there are hundreds of data points and you don't know why they are being used, it's much harder to pinpoint what the axis of discrimination is. Then you add in that these data points are constantly changing over time…'
Equality Task Force member

However, this is not likely to be the main form of unequal treatment that is relevant to data-driven technologies. The rule against direct discrimination insists on strict equal treatment but no more. One of the strengths of algorithms is that they treat all data points 'equally'. But this is also precisely why they tend to project existing inequalities of perception, outcome and experience into the future. It is the very 'neutrality' of algorithms – which treat all data points the same – that reinforces and underscores existing inequality.

The Networkz case offers a good example of the limits of direct discrimination in this context. In Networkz's case, the model not only learned, but revealed, the fact that past patterns of click behaviour correlate with users' gender whether or not gender was included in the machine learning models. Women, in the case study, were more likely to click on lower-paid jobs because gender shapes, and therefore correlates with, people's online behaviour.[107] Networkz could then have decided that it would treat women differently, on the basis of what it had learned: it could offer them only lower-paid jobs. That would be direct discrimination. But the key insight from this case study is that, once Networkz understood how and why this disparity had emerged, it had the means to make adjustments for it and to proactively militate against the unequal outcomes and behaviours that it had observed.

Indirect discrimination

The second form of discrimination prohibited by the EA is indirect discrimination. The EA defines indirect discrimination as follows:

(1) A person (A) discriminates against another (B) if A applies to B a provision, criterion or practice which is discriminatory in relation to a relevant protected characteristic of B's.

(2) For the purposes of subsection (1), a provision, criterion or practice is discriminatory in relation to a relevant protected characteristic of B's if –
  (a) A applies, or would apply, it to persons with whom B does not share the characteristic,
  (b) it puts, or would put, persons with whom B shares the characteristic at a *particular disadvantage* when compared with persons with whom B does not share it,
  (c) it puts, or would put, B at that disadvantage, and A cannot show it to be a *proportionate means* of achieving a *legitimate aim*.

Indirect discrimination looks *beyond formal equality* (like cases being treated alike) towards a more *substantive equality of results*: recognising that criteria which appear neutral can have a disproportionately adverse impact on people who share a protected characteristic, and this hits the claimant in a way that is unjustified.[108]

Awareness of indirect discrimination plays an important role in informing the design and use of data-driven technologies. Without overtly treating individuals differently – indeed often because the designers take care to ensure that historic data are treated 'equally' – the result may be that historic inequalities are compounded and reinforced.

Together, direct and indirect discrimination provide a framework within which to think about the underlying duty[109] to secure substantive equality and equal enjoyment of underlying rights in society.

Applying direct and indirect discrimination law

Networkz's job advert recommendation system, which predicts the probability someone will click on and apply to job adverts based on past click and application behaviour, establishes a potential problem concerning the prohibition of direct discrimination. Was the reason a person did not see an advertisement 'because' of a single, identifiable protected characteristic?

Where datasets with hundreds of variables are involved, and a system is continually learning about accurate predications based on a host of features and their changing relationships that may, or may not, be proxies for a protected characteristics, it may be impossible to know, let alone demonstrate, that those systems directly discriminate against individuals on the ground of a single, self-contained protected characteristic.[110]

Our Networkz case also illustrates the importance and complications of indirect discrimination. Patterns of user click and application behaviour are correlated with both gender and the average income attached to job adverts. Provided the model is accurate and well-calibrated, the model replicates patterns of inequality because of "hidden barriers." But it is not immediately clear whether this falls within our existing statutory framework of 'indirect discrimination'.
The next section considers this in more detail.

Provision, criterion or practice

The object of prohibitions against indirect discrimination is a "provision, criterion, or practice". PCPs have very broad meaning referring to almost any rule, practice, requirement, condition, or criterion which puts, or may put, someone at an unfair disadvantage, which means there are multiple levels of analysis in machine learning to which a PCP could apply. For instance, in Networkz's case, the choice to predict the probability someone will click on and apply to a job could constitute a PCP. Brill's entire XFN decision-making system could be treated as a PCP whose impact analysed against indirect discrimination standards. Alternatively, discrete components of XFN could be identified as PCPs.

Here, the challenge is not the theoretical establishment of a PCP, which should be straightforward, but the absence of any requirement to document relevant practice and choice, or means to access information about these practices and choices across multiple stages of decision-making and analysis. PCPs are being used which no-one has identified, addressed or explained, and which are not transparent to people affected by them.

If multiple data points and criteria are used to make predictions, it may be hard to establish which criterion or criteria have or may have had an adverse effect on members of a particular group.

Establishing disadvantage

To establish a prima facie case of indirect discrimination, a claimant must show the PCP puts, or would put, them and group with whom they share a protected characteristic at a "particular disadvantage" compared with others.[111] The nature of evidence which may be used to demonstrate a particular disadvantage is case-specific, so courts will make context-sensitive judgements about particular cases.[112] But the burden of proving disadvantage rests on the person trying to allege discrimination.

As we discuss above, if multiple data points and criteria are used to make predictions, it may be hard to establish which criterion or criteria have or may have had an adverse effect on members of a particular group. It may be too obscure or too complicated to prove that they are likely to place any particular potential claimant at a significant disadvantage.

Machine learning and other statistical decision-making systems tend to make it *easier* to record and report prima facie evidence of disparate impact or patterns of inequality. In Networkz, as a fast-paced, automated system driven by the use of data, it would be relatively easy for Networkz to record patterns in the average income of job adverts shown to different users. But in practice, Networkz have few incentives to record and report that evidence concerning unequal outcomes because there is no absence of express reporting requirements beyond pay gap monitoring.[113]

Proportionality and justification

Even if it is shown that a criterion has a disparate adverse effect on members of a particular group, it may be justified if it is proportionate to the aims of the business. For example, an engineering firm may be justified in seeking graduates in engineering, even if they are disproportionately male; but a company with a large HR department may not be justified in choosing a course of action involving a disparate impact affecting a small handful of people, although that course is proportionate.

Use of a PCP is proportionate only if it:

(a) corresponds to a *real business need* on the part of the employer, in the sense of being sufficiently important to justify the limitation of a protected right

(b) is an *appropriate means* of achieving the objectives pursued in the sense of being rationally connected to the objective; and

(c) is *necessary* that end, in the sense of balancing the interests of the person using the criterion with the adverse affect of its use on persons affected.[114]

Once a prima facie case of indirect discrimination has been established, the defendant accused of indirect discrimination then has the opportunity to objectively justify the PCP as a proportionate means of achieving a legitimate aim. This involves a retrospective assessment which evaluates and balances the severity of the measure's effects on the relevant groups against the importance of the objective.[115] It covers both procedural and substantive impacts.[116]

The employer does not have to show that it had no alternative course of action to achieve a legitimate aim, but its actions will not be considered 'reasonably necessary' if the employer could have used a less or non-discriminatory means to achieve the same objective.[117]

The concepts of proportionality and justification invite unpacking and scrutiny of the *nature* of the adverse impacts of data-driven technologies; some degree of *comparative evaluation* of these impacts; the *auditing* tools and methods used, and the approach taken to seek *a non or less discriminatory* means to do the job. Where there is clear predictive validity of a reasonable, alternative, less or non-discriminatory means of achieving the outcome sought, a designer or employer should ordinarily be required to select that means. And where there is incomplete or inadequate material to make this assessment, it should be transparent *why* that is, and whether or not that opacity is something that could or should have been corrected or modified.[118] These factors are especially relevant to justification, which is likely to become the real battleground in test cases, although they will assist assessment of proportionality too.

This will require deploying a dynamic range of tools to secure transparency, starting with consistent ways to summarise the patterns and structures in the datasets on which statistical tools are trained so that courts and regulators can understand and interrogate them. Courts would also need a clear statement of the purpose of designing and deploying a tool, the choices made as part of the design process, what comparisons of alternative procedures were made, and how that process is compliant with legal obligations. Developing consistent ways to ensure organisations gather and report that evidence will be critical.

It follows that auditing and equality 'impact assessments' would play a valuable role in supporting organisations develop the tools and capabilities needed to perform and demonstrate this evaluation exercise. This demands some consideration of the different auditing tools and methodologies available, as we have analysed previously.[119] Our research flags a concerning absence of attention given by engineers and others to different approaches and choices that can be made to address or mitigate the adverse impacts of data-driven technologies.[120] We return to this in Part 4: EIAs could be undertaken as part of the a wider algorithmic impact assessment or DPIA, if it is disclosed.

We think these action points should be undertaken by responsible companies straight away. This would also demonstrate compliance and help establish justification if a prima facie case of indirect discrimination is established. But there are limits to retrospective evaluation, undertaken solely as a component of identifying discrimination, in the absence of clear auditing and reporting obligations. And because engineers require specific instructions in order to design legally 'compliant' technologies, muddy water is likely to compound the problems identified in our case studies. Practitioners and academics have highlighted increasing levels of concern among designers[121] about how to respond to these worries, which have not been met by the strong body of work on technical bias.

It follows that using an individual tort model is inadequate, as an individual claimant may not have access to the information necessary to formulate a claim. It is expensive in time and money for a person to bring a claim, particularly a claim which would require costly expert evidence about AI and ML in order to unpack the data and underlying design decisions that have driven a disparity. And, if that is done, it may be just as hard for an employer to demonstrate justification.

So, there needs to be a shift away from individuals being expected to prove that they have been discriminated against, to a model which requires builders and users of these systems to proactively audit and identify potentially adverse impacts from use of that technology, and make decisions about data points, features and deployment which have regard to these impacts. Human decision-makers must be able to demonstrate that the decisions they have made are appropriate, fair and necessary. This evaluation process cannot properly be coded into analytic tools or 'automated' by an ML system. These are decisions which require contextual, qualitative human judgements which must be made, and recorded for evaluation at a later date.[122]

AI ethics principles, as some prominent lawyers have argued, are inadequate to inform standards for this process of evaluation. If ethics principles alone were used to inform this balancing exercise, it would not be difficult for Networkz to argue that p(click) gives people opportunities to apply for jobs to which they might not otherwise be exposed, which might be said to improve collective and individual well-being, despite the apparently mundane commercial context. We return to the specific legal obligations required to enforce equality norms in Part 4.

Positive duties: duty to make adjustments and public sector equality duty

Our review of the existing regulatory landscape has led us to conclude that the duties not to discriminate directly or indirectly are not sufficient to remedy the tendencies of data-driven technologies to reinforce and reproduce unequal outcomes. Individuals who design data systems should be accountable in positive terms for the steps they take, which shape the outputs of those systems. The existing legal framework imposes two such positive duties: the duty to make reasonable adjustments and the public sector equality duty. Both those positive duties have limitations but they offer valuable tools that might be brought to bear in relation to data-driven technologies.

The Thor case study offers a useful access point to the employer duty to make reasonable adjustments for disability. Section 20 of the Equality Act imposes a duty to make reasonable adjustments where a workplace practice or feature puts a disabled worker at a disadvantage. Absent human agency, it is doubtful that a ML system could ever comply with that duty when it does not know what it is 'adjusting' for in the case of an individual.

The duty does not arise if the employer does not know, or could not reasonably be expected to know, that a disabled person is an applicant for a job, or is likely to be placed at a substantial disadvantage by the process.[123] According to the Code, this means that an employer "must do all they can reasonably be expected to do to find out whether a worker has a disability."[124] This suggests that the employer should offer some means to access a human decision-maker from the outset. The public sector equality duty ('PSED') offers another example of a positive duty set out in the existing statutory framework.

It does not apply to Thor, as a private body, and it is structured around the protected characteristics and groups as they are currently identified in the Act. But it deserves particular attention as a model for informing our policy response to the challenges we have identified above. It is a specific primary duty on public authorities to give 'due regard' to identified equality needs, supported by specific duties imposed by secondary legislation (which differ in the different jurisdictions of the UK).[125] The duty only applies to public bodies, in the performance of their public functions, and is only a duty of consideration, not a duty to act. But it does require public bodies to have regard to the need to advance equality of opportunity and foster good relations between people who share a protected characteristic and those who do not.

The PSED has been criticised for the fact that it does not require any action to be taken to address inequality,[126] which limits the way in which the duty can inform strategic decision making. But it does require users to identify – and squarely consider – the unfair and unjustified equality impacts which may arise or compound as a result of their decisions. So the PSED offers a useful model of awareness raising. As discussed above, one of the challenges in this area is that designers, coders and data analysts may consider that their work is 'neutral': the data is merely processed in a 'fair' manner and questions of equality have little or no relevance in the machine learning space. Awareness is a necessary first step before accountability. The PSED will already apply to a small number of mass data decision makers (though we doubt that they apply it in the manner we have outlined above). We consider that the duty to have regard to equality impacts should apply much more broadly – in both the public and the private sector – to those who design and use mass data systems. We return to this theme below, when setting out our proposed new statutory framework.

4

Key challenges and gaps

Part 4

Key challenges and gaps

This section outlines the central challenges and gaps identified by applying the law we have summarised in Part 2 to our case studies. We summarise these challenges under three categories: the content of EA legal obligations; their scope; and their enforcement and realisation in practice.

Content of legal obligations for equality

Limitations to current approaches to 'equal treatment'

As outlined in the first section, both complex machine learning models and simple algorithms like Ofqual's reproduce past patterns of inequality. When datasets capture patterns of inequality and disadvantage – such as disparities in educational attainment in schools across the UK – the predictions of statistical tools trained on those data will reflect those patterns of inequality and disadvantage without intervention.

'Absolutely the problem with anti-discrimination law is that the onus is focused on an individual, and it is always after the fact.'
Equality Task Force member

A statistical model learns to treat people differently based on statistical differences observed with respect to the specific 'prediction' task at hand. Networkz's model predicted that men and women would click on job adverts with different average incomes because men and women did click on job adverts with different average incomes. Yet when predictions that reflect disparities across social groups are used to make decisions, those decisions compound that very disparate impact.[127]

There is no neutral or fail-safe way to build data-driven decision-making tools, because even accurate and unbiased statistical models will reproduce existing patterns of inequality and disadvantage. If we are to ensure these tools do not compound inequality, they must be deliberately designed and deployed to advance and promote inequality across the entire innovation cycle.

This suggests that narrow, formalistic understandings of equality of treatment will not achieve the underlying purpose of the Equality Act: to improve outcomes for those who experience disadvantage. Insisting that data-driven systems be "blind" to categories of disadvantage will simply ensure those systems

reproduce existing patterns of disadvantage, whether across protected characteristics like race or gender, or other characteristics like geography, socioeconomic status, or the school someone attended.

The Networkz case makes this clear. "Blinding" the model that predicts click probability to gender, for instance by removing gender as an input, does not prevent the model from replicating patterns of gender inequality encoded in the data. In fact, the best way to address those patterns may be to *include* gender as an input into the model, because it enables the model to make more fine-grained predictions in full knowledge of existing patterns of disadvantage.

In building and using data-driven technologies, there is no neutral option. Aiming to build tools that are simply "non-racist" or "non-sexist" will ensure that those tools replicate and reinforce existing and enduring patterns of racism and sexism. By contrast, building statistical tools to be anti-racist and anti-sexist will often require deliberate consideration of race and sex when building decision-making systems.[128]

The Task Force believes that sensitivity and awareness to difference is the best way to interpret and enact principle of equal treatment. Persistent categories like race and gender constitute legitimate grounds on which to treat people differently *for the purpose* of ameliorating and mitigating that disadvantage, provided the means of addressing disadvantage is proportionate.[129] When persistent patterns of inequality are encoded in data, we need to look at the nature and cause of inequalities which are relevant to a decision so as to properly inform our choices, and to take reasonable steps correct unfair disadvantages.[130]

As we argue in the next section, if preventing data-driven technologies from compounding inequality requires them to be deliberately built to advance equality, then law and other regulation may need to explicitly impose broader and more robust positive duties to advance equality, which cannot be constrained by formalistic understandings of equal treatment that may inhibit the capacity of organisations to address the disadvantage that they have identified.

Individual rights and remedies are not sufficient

As statistical tools replicate existing patterns of inequality, encoded in data, those who are disadvantaged are disproportionately impacted by the use of patterns in data to make decisions. If gender tends to correlate with clicking on lower-paid job adverts, then predictions of click probability will have disparate impact on women. The Ofqual case made this clear to the public: using data about past disparities between predicted and attained grades may disproportionately burden students from disadvantaged backgrounds, not because the algorithm was biased per se, but because those students are generally disproportionately disadvantaged by the quality of their educational environment.

Our analysis has shown that statistical decision-making systems make assumptions to predict the future behaviour of individuals based on group or community stereotypes.[131] This means that, to understand impacts, we must examine group and relative outcomes. The replication of existing patterns of disadvantage by statistical tools will reflect and reinforce disparities between demographic and social groups, as well as other enduring dimensions of inequality. In this context, individual rights and remedies alone are an inadequate tool for ensuring data-driven tools are built to promote equality.

First, it will be extremely difficult to detect systemic-level impacts of statistical systems within isolated, individual cases. Law must structure relationships of power between citizens not just as individuals but as members of social groups, especially groups which have historically been subject to injustice.[132] This means we need better information gathering mechanisms to aggregate individual cases and understand harms at the level of social groups.

We have seen that statistical systems and predictions can offer insights into the contours of systemic inequality. But achieving this is not possible within a regulatory system which only makes such insights visible through individual rights claims. Existing access and transparency rights do not permit individuals to access of information about others. This means that *group or relative* outcomes are obscured. The need for group-level governance may be seen in recent ICO guidance tightening up its advice emphasising the need for consultation of data subjects and their representatives.[133]

Second, orienting equality law around individual rights may not be an effective way to ensure statistical tools are built in ways that prevent inequalities of power from being compounded and becoming entrenched. Power includes the ability to act, or not act, and to influence or control something. It exists in the relationships between people and groups, rather than belonging to one person. Taking a systems view of the equality impacts of data-driven technology, including the increased concentration of power in the hands of a small number of corporates that control much of the design and relevant data sources relevant, we need to look beyond an individual interface with the tool itself.[134] So while individual rights are an important safeguard against individual harms, those rights are often not adequate, and sometimes a distraction from, structuring accountability over the organisations and bodies that exercise power over citizens.[135]

This is why our policy recommendations focus directly on structuring accountability rather than simply extending individual rights.

There are multiple forms of unfair discrimination

Data-driven decision-making technologies discriminate by definition – their purpose is to differentiate between individuals based on characteristics shared with others in order to accurately predict some outcome. Some of these may be innocuous, for instance if statistical tools accurately predict what kinds of raincoat or food people prefer. But many forms of statistical discrimination may constitute unfair and illegal discrimination – and not all of these may be adequately described in terms of protected characteristics.[136]

The EA is built on the concept of protected characteristics as forms of discrimination are understood in terms of protected groups who share a protected characteristic. While protected characteristics have historically been a potent tool for addressing disadvantage, since so much of the disadvantage in our society falls along the lines of protected characteristics, like race and gender, those characteristics can sometimes serve as ineffective proxies for disadvantage.

This is especially true with data-driven technologies. The use of any kind of statistical average to make decisions about individual people will disproportionately burden those who are already disadvantaged, replicating and entrenching past patterns of disadvantage, regardless of whether those patterns correlate with protected characteristics like race and gender.[137] The complexity and range of datasets informing the tools, combined with new technical capabilities, mean that we are seeing 'new' forms of unfair differentiation in addition to well-established ones.[138]

For example, the Thor case study demonstrates how pervasive correlations between socio-economic background and other variables are in datasets, such as postcode, place of birth, use of specific types of language, and tone of voice. Sophisticated data-driven technologies may reveal other unexpected correlations which may also correlate with familiar dimensions of disadvantage, such as the language people use online, the social groups they tend to be a member of, or even the kinds of photos they share online.[139]

It may be increasingly important, therefore, to clearly separate out two kinds of provisions: prohibitions against discrimination, whether direct and indirect, that are grounded in and must be claimed on the basis of protected characteristics; and positive duties to make reasonable adjustments or advance equality, which need not be limited to protected characteristics, but rather, which should be focused on addressing existing dimensions of disadvantage, whenever those happen to be unearthed.

There will be a critical need to knit together accountability mechanisms to enable challenge to new and newly recognised forms of intersectional discrimination, with new rights of access to information, both for individuals and for groups. More generally, as the widespread deployment of data-driven technologies unearths more of the ways in which structural inequalities are connected, we should grasp the opportunity to explore new ways to ensure organisations deliberately promote equality as they build and use statistical tools in decision-making.[140]

Scope of obligations

Obligations of the EA are ex post facto

Because the EA's central provisions oriented around individual rights, they can often be leveraged only after the fact, when claimants bring individual suits alleging they have been discriminated against. Although there are some persuasive scholarly arguments that indirect discrimination in theory entails an ex ante duty to identify and mitigate possible disparate impact, in practice, this is an unreliable model to ensure unfair disadvantage is not replicated *before* a decision-making system is deployed.[141]

Early action is needed because statistical systems operate with unprecedented speed and on an unprecedented scale, raising the stakes of how particular machine learning models are designed, and because decision-making power is diffuse. So, equality duties must commence at the start of the cycle and supply chain of a designing and deploying a system, and must continue throughout its use.[142]

It also means that explicit duties may be needed to ensure that organisations do, in practice, give consideration and evaluation to how best to effectively advance equality, from the outset of their design. Appropriate consideration of the full range of mitigation techniques, alternative means to achieve the business aim, and appropriate ways to prioritise between different types of adverse impacts, is essential and should be clearly prescribed.

Pre-emptive governance will also incentive more systematic and ongoing monitoring and evaluation of equality impacts, rather than retrospective evaluation as and when claimants bring individual discrimination cases. Just as technologies like machine learning can be used to efficiently promote equality instead of compound it, so they can be used to effectively gather and record information about patterns of disparity encoded in datasets and reproduced in decision-making systems.

“

The opposite of ‘racist’ isn’t ‘not racist’, it is ‘anti-racist’.

Ibram X Kendi, writer

EA's positive obligations do not extend to the private sector

We have seen that the EA does contain important existing 'positive' equality duties including the public sector equality duty. But precisely because the human choices that shape the effects of data-driven technologies are distributed across a wide range of roles and organisations rarely confined to the public sector, they cannot be confined to the public sector, or specific commissioned services[143] alone.

Many of the most important choices about the design of statistical systems that shape citizens' lives – whether Networkz's advertising delivery model or Thor's hiring system – are made in private sector organisations. Moreover, these technologies change the point at which humans direct decision-making, requiring a focus not only on individual decisions about individual cases, but on how the tools are integrated into wider decision-making systems.[144]

The traditional grounds for distinguishing between distinct obligations for private and public are neither sharp nor persuasive in their application to the design and deployment of data-driven technologies.[145] So, without extending positive duties and ensuring they catch all actors, including private sector actors, these choices will simply be left to the requirements of negative prohibitions against discrimination. And in the design of statistical tools, without that deliberate intent, those tools are likely to entrench and compound existing patterns of disadvantage.

Even the PSED, bolstered by rule of law principles as they apply to public bodies,[146] is founded on existing axes for discrimination critiqued above, and has been subject to some critique by the EHRC and academics for reliance on tacit knowledge, failure to target the most persistent inequalities and lack of transparency about the specific actions needed to achieve its objectives. Nonetheless, the PSED is a successful model for an affirmative duty which we believe should be strengthened and extended, as we outline in our recommendations below.

While there are mechanisms which lever some degree of transparency as to decisions taken in the design of ML tools used at work in the GDPR, as we have seen, these are inadequate in making transparent outcomes, and patterns of harm. This means that private organisations, in particular, have few incentives, and often lack the capacity, to evaluate, record, and adjust the equality impacts of statistical tools.

The traditional grounds for distinguishing between distinct obligations for private and public are neither sharp nor persuasive in their application to the design and deployment of data-driven technologies.

Part 4
Key challenges and gaps

EA duties do not extend across the entire design cycle

'Who gets to decide what is reasonable in a particular context? Who has the authority to do that?'
Equality Task Force Member

The case studies, supported by our hiring research,[147] demonstrate the foundational importance of human decisions in shaping outcomes of ML tools for equality, at the earliest stage of the design process. They also show how many professional actors, organisations and digital platforms are involved before the stage at which the system is activated by the employer. Our analysis of the application of current law also reveals that its hooks and tools are inadequate to ensure 'equality by design'.

As we highlighted in Part 1 it is the very earliest decisions made in the design and deployment of these tools which shape the purpose, target and operation of the system. Some discreet obligations may apply to those who design or sell a system (for example inducement under s111 Act, or the provision of a service under s 29 Act) but these are isolated and do not extend to all the decision-making roles we have identified in this report.

IFOW workshops with the IET indicate that there are high levels of uncertainty and clear need for direction about responsibilities across the technology cycle. An explicit principle of equality by design, with attendant obligations, would encourage reflection and attentiveness to the issues we have identified throughout the design process. This would help ensure that all those involved have more than partial knowledge of how the system functions as a whole, mitigating the 'black box problem' for end user employers, as well as employees, who may not know or have any direct contact with engineers, manufacturers, platforms or others involved in the supply chain.[148]

Enforcement of obligations

Existing reporting requirements are limited

The widespread use of data-driven technologies could be used by organisations to monitor equality outcomes with limited administration. Equally, such data could provide ready evidence of prima facie indirect discrimination. For this, such statistical information must be readily accessible.

We have seen that there are at present few incentives for businesses to record, evaluate and report those patterns, as the only enforced equality reporting obligation requires medium- and large companies to report gender pay gaps. Unless reporting requirements are extended to a wider range of organisations, sensitive characteristics, and sectors, there will be few mechanisms for citizens, strategic litigation bodies or regulators to leverage the opportunities offered by data-driven technology to detect inequalities and interrogate the decision-making systems that reproduce them. The lack of transparency is also hindrance to wider application of the current law.

It is misguided to think that information essential to deciphering accountability is hidden within an inaccessible 'black box'. Human choices set the defined outcome, set the variables and ultimately are responsible for overseeing the patterns resulting from the model once it is trained on data. In turn, processes which document these decisions as they are taken are critical to structuring accountability for them effectively. Shifting the burden of proof will help, but this is a procedural safeguard, rather than a substantive solution, and will only bite during the process of litigation.

Equality impact assessments may be one such tool to support evaluation of outcomes, if undertaken rigorously and on an ongoing basis, so that adjustments can be made, as we have previously proposed.[149] Our public consultation suggests support across business, academia and unions for a requirement which recognises the practical benefits of an early, focused evaluation of equality impacts. Mandatory annual reporting obligations, attached to EIAs, could thread together the different accountability needs we have identified.

Models of managing intelligent and responsible design may also take other forms. As can be seen in the governance of our built and natural, rather than digital, environment, a range of professionals subject to different professional certifications, democratic expectations and collaborative procedures, can support shaping outcomes in place. Those deciphering an appropriate form for governing our data-based architecture should learn from these.[150]

Equality regulators lack powers and resources

This report has argued there are significant gaps in the mandates and resources of existing regulators. The UK's regulatory environment is made up of multiple actors, enforcement agencies, inspectorates and ombudsmen with a range of responsibilities (see Annex 1). Some regulators have an explicit remit to consider or address bias and discrimination but this is not uniform. This means there is a mixed picture of both responsibility and accountability.

Transparency mechanisms are limited and not effectively deployed or enforced; regulators with overlapping mandates for enforcing equality obligations are not collaborating to the extent required; existing obligations on the private sector are not effectively monitored and the scope of those obligations is too narrow; and the Equality and Human Rights Commission (EHRC) in particular has neither the social or material resources, nor the mandate, necessary to enforce compliance with equality obligations in an age of ubiquitous data-driven technologies.

What's more, some government offices, regulators and government advisory bodies tend to minimise, rather than prioritise equality policy, so that it is seen as a secondary consideration or afterthought, rather than central and cross-cutting consideration. For example, the CDEI Bias Review did not have a forum for formal input from the EHRC.

To make the UK a hub for *developing* AI we need to build on relative strengths in *governing* it too: we need a new structure and mechanisms for accountability and an office with clear cross-sectoral mandate to support our exiting regulators.

Such a boost in resources, capacities and powers should enable our regulators to develop in-house capabilities to pursue and execute large-scale PED and discrimination investigations and test cases. These will be enormously complicated and sensitive and will require specialist, interdisciplinary teams. In addition, investment in capacity should provide for technology expertise including secondments from industry and academia. Funding should be earmarked for legal costs and complaints from multiple individuals affected by the same algorithmic system should be allowed. EHRC monitoring and enforcement action should not be delegated to for-profit companies or third-sector institutions.

To make the UK a hub for *developing* AI we need to build on relative strengths in *governing* it too: we need a new structure and mechanisms for accountability and an office with clear cross-sectoral mandate to support our exiting regulators.

There is a striking disconnect between legal regimes and regulators

We have seen that understanding of the issues, and the levers within and between our existing legal frameworks is fragmented. Even more importantly, in widespread acknowledgement that cross-regulatory working is required, we have observed that this has not happened in practice. The ETF, for example, hosted the first direct and AI-specific dialogue on some intersections between data protection and equality law by convening representatives of the ICO and EHRC. This reflects the siloed responsibilities of the relevant institutions as identified in Part 3.

The Government has recognised that more cohesive regulatory approaches are needed and established a new 'Digital Regulation Cooperation Forum', bringing together the CMA, ICO and OfCom. This seeks to enable coherent, informed and responsive regulation of the UK digital economy which serves citizens and consumers and enhances the global impact and position of the UK.[151]

To address the challenges posed to work by AI, such a knitting together of powers, knowledge and skillsets is needed. The joint investigations and strategic test cases will need close cross-regulator collaboration, and the development of a joint statutory code on the application of existing requirements. However, as we have discussed through this report, this is only part of the answer. Further clarification, development and scripting of law is necessary to address the fundamental problems and accountability gaps we have identified and create a framework for change.

5

A new path forward: an Accountability for Algorithms Act

Part 5

A new path forward: an Accountability for Algorithms Act

There is nothing inevitable about the way data-driven technologies shape our future of work or our lives.

Humans make technology, the design and deployment of which cannot be neutral, as we describe in Part 1. The institutional context raises some distinct problems for accountability, which we examine in Part 2, and our current regulatory ecosystem is strained. Part 3 and 4 identify gaps and inadequacies in our approach and frameworks for accountability which are prohibiting the ability of our regulators to work together and address the challenges we have identified.

These challenges sit at the interface between data protection and equality law, hiding behind the myth of neutrality, and inhibiting technology and regulation alike from serving the public interest.

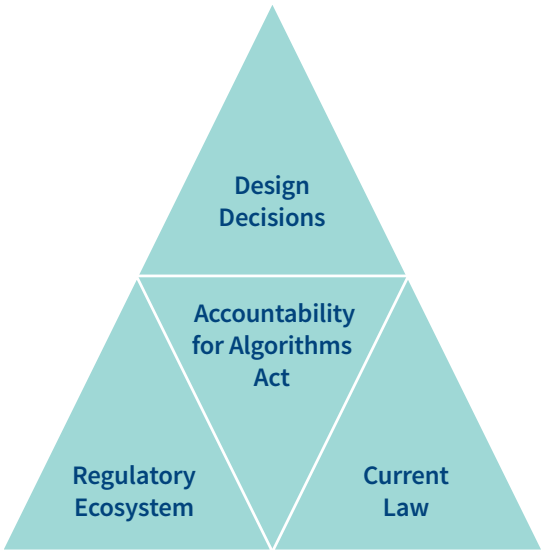
Our policy response to these challenges must be a human one too. Law is a signifier. Initiated and shaped by humans, its ambition should be to change human and organisational behaviours in ways which serve the public good. Law can accomplish its goals directly, but it can also change priorities and attitudes towards the regulated behaviours. It should reflect social and ethical norms, but steps up, when these norms are not producing the actions and behaviours which are needed. But the law has been outpaced.

Each part of our analysis has taken us to the need for an overarching legal framework to reaffirm individual and collective human agency over, and accountability for, algorithms. At the centre of this new legal framework should be new legal duties, and means of individual and collective redress, which respond to the fundamental and multi-faceted challenges we have identified: in particular that data driven-technologies will compound and project *different forms* of individual *and collective* inequalities *into the future* without intervention. As the Ofqual case has highlighted, these are new forms of collective harm which have, so far, escaped mainstream attention of the public and policy-makers. We think this must change.

So, we need a new approach to governance and regulation of algorithms, including AI and ML. This approach must be principle-driven and human-centred, work across the entire innovation and deployment cycle, shift our emphasis to preventative action, and align our legal regimes and regulators.

Part 5
A new path forward: an Accountability for Algorithms Act

A new Accountability for Algorithms Act



This joined up approach must provide clear direction: the principle of ethico-legal equality between citizens and social groups must be a central pillar of the new regime, and not an afterthought. Unless technology is *deliberately* built to advance this principle, our analysis has shown that they *will* compound structural inequalities. Our proposed regime therefore repositions the principle of equality, and overarching objective of equality law, to underpin the new regime for algorithmic accountability: to secure equal enjoyment of fair opportunities.

This means that like cases must be treated alike (unless there is a lawful and proportionate reason for not doing so) and that unlike cases must be treated differently (so that an equal rule with unequal effects must have proportionate justification). We think that enduring structures of accountability should be founded on this principle and ensure that individuals and organisations from designers, platforms, private and public employers are held to account for decisions which do not comply with it, together with the more established principles in AI governance, including fairness, transparency, sustainability, safety and privacy.[152]

Our focus has been algorithmic machine-based decision-making at work. But we are mindful that many of the issues we have examined in the context of work extend beyond this, and regulatory response cannot be limited to it. The decision-making we have looked at under the magnifying glass here, which determines access to work, or fundamental terms of it, may be a 'bellwether' for wider challenges. To maximise the potential of data-driven technology, spread its undoubted benefits and build public trust in its use these must be addressed. We think our proposal may also contribute to wider national and international efforts to regulate artificial intelligence and machine learning which are gaining traction, and may help the UK provide a leadership role in this regard.[153]

An Accountability for Algorithms Act

We propose an Accountability for Algorithms Act ('AAA') to: provide clear direction across the different actors involved through the technology life cycle; fill the gaps we have identified; and provide a shared mission to unite our regulators. The AAA would regulate significant algorithmically-assisted decision-making, which meet a risk-based threshold, across the innovation cycle, legal spheres and operational domains in the public interest.

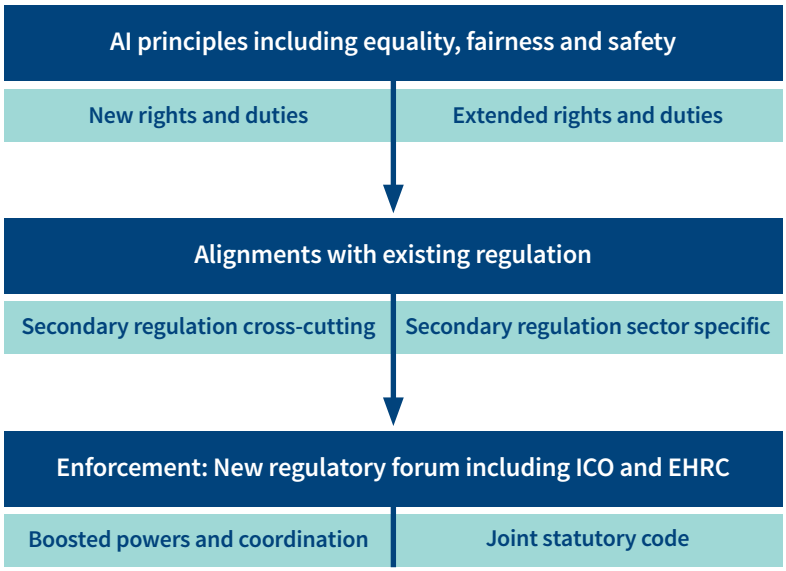
Drawing from the Data Protection Act, Health and Safety and Work Act and Environmental Protection Act, the AAA would be an umbrella, 'hybrid Act', combining overarching principles, to give well-established norms in AI governance a statutory base, with new duties, and standards for 'self' regulation to allow for a fast-changing landscape.

The AAA would provide for joint ICO-EHRC statutory guidance, and detailed sector-specific guidance under secondary regulation, and it would amend existing legislation as appropriate, to ensure consistency. The AAA would also establish a new regulatory forum and powers to support access to justice and

enforcement. More work is needed to consult on and develop the proposal, including designing some appropriate caveats for intellectual property and national security. This should start as soon as possible. We also note that extra-territorial application and competition policy will need particular attention in developing a comprehensive framework for accountability for digital services and actors. As the new US Judiciary Committee report argued,[154] the underlying purpose of competition law and anti-trust is also to hold corporations accountable, and to structure market power, in the public interest. But this is outside the scope of the ETF report.

We propose using the concept and language of 'reasonable adjustments' to define the new duties which are central to the AAA. We anticipate that a new body of precedent and law of the duties of the AAA would quickly build up applying general principles to specific categories of cases, in a similar way to that developed by the Financial Ombudsman.[155]

A new Accountability for Algorithms Act



New duties: prior evaluation and adjustment

The outstanding feature of this Act would be new corporate duties of prior evaluation and reasonable adjustment, making a shift in regulatory emphasis to pre-emptive governance and action in the public interest. Our focus is equality impacts, but our duties recognise that there are other adverse impacts, also capable of collective harms which will need consideration. As we have established, expecting individuals to identify 'torts' and reliance on retrospective evaluation by a judge is unsatisfactory model for engineers, employers, and employees alike. Our proposed new duties recognise this, and respond to the core challenges which run through our report.

This proposal takes into account IFOW's equality impact assessment prototype, published in April 2020 as a proposed voluntary self-assessment measure. IFOW ran a public consultation in which a range of stakeholders acknowledged the value of prior evaluation and reasonable adjustment[156] of equality impacts, and the need for a 'harder' approach to achieve this. Equality impact monitoring must be ongoing, based on sound evidence, high quality analysis, and the development of capabilities to make appropriate adjustments. We anticipate this would be an iterative process, with standards developed and raised by the regulators over time.

Part 5
A new path forward: an Accountability for Algorithms Act

New statutory duties for public consultation

- **Duty on actors who are developing and/or deploying algorithms, as well as other key actors across the design cycle and supply chain, to undertake an *algorithmic impact assessment, including an evaluation of equality impacts*, or a dedicated equality impact assessment.**

  This duty would be subject to a risk-based contextual threshold which will be developed. Our primary concern is the use of algorithmic systems to determine access, terms or conditions of work.

  This assessment should be rigourous, dynamic and ongoing through the design life cycle and deployment of the system. It would be be supported by a statutory code which would set out factors to be considered but would not prescribe a fixed framework for the evaluation process.

- **Duty upon actors who are developing and/or deploying algorithmic systems, as well as other key actors across the design cycle and supply chain to make adjustments which are *reasonable in the circumstances of the case*, with regard to the results of the equality impact assessment.**

  The purpose of this duty is to eliminate unlawful discrimination and advance equality of opportunity and fair outcomes between people and groups who share a protected characteristic, and those who do not, in decisions taken with the assistance of data-driven technology.

  This duty would require consideration of key factors identified in the AAA (including the equality impact assessment, cost, nature and extent of disadvantage, reasonable alternatives and compliance with other duties) but not prescribe an approach to determining what is reasonable. The duty would cover public and private actors.

  The new duties, new regulatory forum and its joint investigation and test cases, with the statutory code, would result in a new body of common law which would develop to provide closer guidance on what amounted to 'reasonable' adjustments in a given case.

- **Duty for actors across the design cycle and supply chain to co-operate in order to give effect to these duties.**

  Some level of co-operation, communication, and disclosure would be needed to give effect to the primary duties in this section. The statutory code would provide further guidance.

- **Duty to have regard, while making strategic decisions, to the desirability of reducing inequalities of outcome resulting from socio-economic and also place-based ('postcode') disadvantage.**

  This new duty builds on s 1 Equality Act which is a public sector duty regarding socio-economic inequalities by extending it to (a) private sector and (b) place-based disadvantage. The duty is aimed at reducing *inequalities of outcome* and is not dependent on identifying or establishing a particular protected characteristic.

Part 5
A new path forward: an Accountability for Algorithms Act

Increasing transparency

Throughout our analysis, we have highlighted areas in which actors would benefit from increased transparency about the nature and roles of human-decision-makers as well as aspects of the algorithmic systems themselves. We think increased transparency about key decisions across the innovation cycle and supply chain would plainly benefit most actors, and certainly benefits those at the receiving end of the decisions at hand.

Each of these proposed duties need public dialogue, further consultation and development.

Innovation cycle transparency

- **New mandatory transparency obligation to record and report on *facts of, purposes and outcomes* of algorithmically-assisted decision-making, subject to the risk-based threshold.**

  This duty would be mirrored by a new 'duty to know' about algorithmically-assisted decision-making, subject to the risk-based threshold.

  The AAA statutory code will specify minimum standards.

- **Duty to record and publicly disclose a summary algorithmic impact assessment (AIA), including the assessment of equality impacts, or the dedicated equality impact assessment (EIA).**

  Individuals and collectives should be entitled to receive additional information on request. Regulators will be entitled to the completed AIA and EIA, including choice and evaluation of training data sets and code on request. AAA statutory code will specify minimum standards.

- **A new right to know. This right would include access information about performance of the new transparency duties i.e. the fact of use, purposes and outcomes of algorithmically-assisted decision-making above the risk-based threshold; and the summary AIA, including the assessment of equality impacts (or the dedicated EIA).**

  A freestanding right to an explanation would be extended beyond technical design decisions to the human decisions we have identified including the purpose, capabilities and limitations, remit, model, logic involved; plus basic information about the training datasets (their perimeters and metadata) and methodologies, processes and techniques used to build, test and validate the system; and the human roles and oversight. The Turing/ICO guidance on 6 types of explanation may provide a sensible basis for the right to an explanation but its precise remit should be subject to further consultation.

Part 5
A new path forward: an Accountability for Algorithms Act

Support for collective accountability

Although the focus of the AAA is on new corporate accountability duties and transparency, we recognise the important 'mirror' role of individual and collective rights for their own sake and as a means of enforcement. There is no inherent conflict between group and individual rights, or between group or individual fairness measures.[157] To the contrary, group fairness is necessary and will enable fairness on an individual basis.

So the AAA aims to boost support and access for unions, as our national institutions representing worker groups, and create structures for accountability in which individual and collective mechanisms for accountability are harmonised. These recommendations build on the growing success of the social partnership model through the pandemic.

Each of these proposed duties need public dialogue, further consultation and development.

Collective access

- **The new right to know and expanded right to an explanation would be excercisable by unions, other collectives and NGOs in prescribed circumstances, with the permission of individuals.**[158]

  This means that employers and other end-users will themselves need to access information from other actors in the design cycle or supply chain about any of the key decision-making points/information relevant to their own obligation in order to provide information to workers or collectives exercising their right to an explanation. It also means that employers should inform relevant trade unions when algorithmic systems are used to determine access, terms or conditions of work.

- **Right for workers to *be involved* to a reasonable level in the development and application of algorithmic systems involving AI used at work.**

  This would extend the application of existing consultation laws, and would be exercisable by unions on members behalf.

- **Digital access to all members and potential members to unions.**

  All workers should be able to *access union support in order* to enforce these new rights, expressly including use of algorithmic systems involving AI.

  This right would mean that workers across different contract types, sectors and the gig economy would be entitled to access union membership.

- **Union equality representatives would work on a statutory footing, with specific data access entitlements.**

  Data access requirements would be extended to all union representatives.

Part 5
A new path forward: an Accountability for Algorithms Act

Clarification: overlapping regimes

Existing laws with implications for accountability would be amended as appropriate to be consistent with the purpose and provisions of the AAA. These amendments are essentially procedural.

Amendments

- Suspension of the PSED duty for Covid-19 should now be lifted.
- The Equality Act and Data Protection Acts would be amended to align with the AAA, including general exceptions in both Acts as appropriate.
- The EA should be amended to allow for claims involving intersectional discrimination. The restriction on identification of a single protected characteristic must be lifted forthwith and by March 2021.
- S1 Equality Act (duty regarding socio-economic inequalities) should come into force in England at the same time as the duty commences in Wales (March 2021).
- The Companies Act would be amended to require Directors to give due regard to equality impacts when making decisions about the introduction of technology.
- The restrictions and safeguards in Article 22 (regarding solely automated decision-making) would be extended to hybrid and semi-automated decision-making decisions. The EA would be amended to clarify this.
- To affirm human agency, humans must always be clearly 'in command' when decisions concern access to work, fundamental terms and conditions of work involve algorithmic decision-making. This means that strategic decisions, with potential for collective impacts on equality must not be 'automated' and individual decision-makers must be identified.
- The Questionnaire Procedure should be bolstered and reintroduced, enabling individual and group applicants to access information relevant to duties in the AAA.
- The burden of proof of discrimination in the Equality Act should be reversed.
- Tribunals would be allowed to entertain individual claims under the AAA, and to make recommendations in relation to the workforce generally.

Part 5
A new path forward: an Accountability for Algorithms Act

New regulator forum

Whilst we do not think there is need for a new regulator, the AAA would establish an intersectional regulatory forum to coordinate, drive and align the work of our regulators, and enforce our new duties, which would otherwise lie between the EHRC and ICO. New powers would be needed to support specialist joint investigations and test cases, improve access to justice, and provide for cross-cutting statutory guidance, which would be combined detailed, sector specific guidance by others. This approach builds on the Regulators Code which requires cooperation between regulators in principle.[159]

To do this, and support closer enforcement of existing duties under each regime, a significant boost of resources is needed. The EHRC, in particular, must not be scapegoated or required to produce complex intersectional guidance on the multi-dimensional accountability challenges we have identified in this report without the means do so.

Each of these proposals need public dialogue, further consultation and development.

Coordinate, drive and align

- **A new regulatory forum including the ICO and EHRC would be established and appropriately resourced to monitor and enforce duties and rights under the AAA.**
- **Both regulators would receive significant funding boosts and support for specialist academic and industry secondments independently, in addition to funding for the forum.**
- **Joint statutory guidance would be issued on all matters covered by the AAA.** The forum would determine which bodies to seek help from in order to devise the guidance, for example from the Turing Institute or CDEI. The statutory guidance would establish a set of standard evaluation metrics and guarantees that actors who put algorithmic systems on the market must provide to their customers, which cover certain basic elements of algorithm performance that are pertinent to equality considerations.
- **Regulators would be entitled, on request, to the complete algorithmic and equality impact assessments in full, together with all relevant sources codes, including training datasets, methodologies, processes and techniques.** This new power would be exercisable where information may be relevant to an investigation or test case. Including code, training datasets, methodologies, processes and techniques.
- In addition, the EHRC and ICO should issue their own statutory and technical guidance on matters relevant to algorithmic accountability which fall within existing remits. In particular the ECHR, appropriately supported, will need to provide guidance on how to demonstrate statistical disparity and assess proportionality in cases of indirect discrimination. And the ICO will need to offer further guidance on collection of sensitive data, to support new and existing equality duties.
- **New powers would allow for suspension of use of algorithms pending investigation or test cases;** and to create or approve of certification schemes involving a set of standard evaluation metrics and guarantees before algorithmic systems are put on the market.

Part 5
A new path forward: an Accountability for Algorithms Act

Changing the design environment

We recognise law is not the only mechanism for change, and will take time to develop. So we have considered a number of other actions which may strengthen the regulatory ecosystem to promote equality in the design process in parallel to the development of the AAA. Many for these will contribute to the development, refinement and implementation of the Act as well.

Proposed legislation always needs very wide consultation. Proposed legislation about the nature, extent and comparison between equality impacts, deciding what fair and reasonable outcomes and adjustments might be for organisations and platforms of different sizes, powers, capabilities and reach, invites social dialogue even more than usual. IFOW have undertaken a discreet public consultation on our EIA, but this is just the start.

The stakeholder and public consultation should include low cost, non-litigious means to enforce new obligations via the regulators and regulatory forum, as well as courts and employment tribunals.

Public dialogue and wide consultation

Public consultation and dialogue on the AAA should be initiated as soon as possible. Areas for consultation should extend beyond the remit of this report.

IFOW target
IFOW will engage in wider consultation about the form of AAA and nature of the duties proposed, with support from ETF members where possible. We will explore others' proposals and improve our own, including new international efforts to regulate AI.

We will co-host a workshop with the Ada Lovelace Institute, which has a remit to convene dialogue and diverse voices on the social and ethical implications of AI and data, in 2020 on the practicalities and challenges of regulatory investigation of algorithmic systems for equalities impact.

Industry Standards

The law provides a base line for conduct but we intend to support raising the 'bar,' as well as raising the floor. Industry standards will important to raise the bar, and feed into the sector specific guidance.

IFOW target
Building on our workshop in October 2019, we will collaborate with Institution of Engineering and Technology and others to initiate dialogue and discussions about possible codes of practice on the design and use of algorithms, AI and ML.

IFOW collaboration
We are co-developing guidance for business with CIPD on responsible use of technology which will extend to equality impacts.

continued

Part 5
A new path forward: an Accountability for Algorithms Act

Increasing Insight into the problem

ETF collaboration
ETF will work to support test cases and joint cases to ICO and EHRC as appropriate.

IFOW target
We will collaborate with Prospect Union to take test data and union subject access requests to employers and the ICO.

IFOW target
We will also encourage systematic mapping of deployment and register of systems involving some uses of algorithmic AI and ML systems to inform the public and policy-making debates, and to support further research by IFOW and others. We are proposing national statistics on some types of technology use to the ONS.

Invitation for collaboration
To seek information and the extent to which informal regulation can address any of the challenges presented, we are looking for firms who want to pilot our proposed EIA.

Worker voice

IFOW pilot
We will use our social policy design methodology to develop new ways to involve workers in the design and implementation of AI and ML technology.

IFOW collaboration
We are co-developing a pack with Prospect for unions to maximise use of existing data protection law.

Raising awareness

ETF target
Members of the ETF will share this report and recommendations as widely as possible, and consider ways to develop them.

IFOW target
Raise awareness of policy-makers about the use of these technologies, and current worker rights, as this limits the extent to which individuals will bring forward challenge, or can provide meaningful consent to share data. Raise public awareness as much as possible, including via other organisations.

“

When there are hundreds of data points and you don’t know why they are being used, it’s much harder to pinpoint what the axis of discrimination is. Then you add in that these data points are constantly changing over time…

Equality Task Force member

End

The purpose of technology governance and regulation must be to enhance human capabilities, agency and dignity and to hold people accountable for the choices that they make. The opportunities these technologies offer are greater than simply enhancing efficiencies, and regulation of them can and must achieve more than simply preventing harm. Data-driven technologies can help break down and expose the connections between patterns of disadvantage just as efficiently as they can reproduce them. Correctly used, they are capable of reducing inequalities and augmenting human skills. But we have seen that they must be deliberately built to achieve these goals, because there is no neutral way to design or deploy data-driven technologies.

Unless we confront the underlying challenges we have identified in this report, and establish a structure for the meaningful accountability of algorithms, unjustified inequalities will compound, and the public trust deficit will be exacerbated. It is for this reason we believe that a precise focus on equality in the governance of data-driven technologies is a matter of public interest. We must now structure accountability of private powers to promote public good, instead of legitimising the collective harms and unfair replication of structural inequalities.

We urge policy-makers to explore how we can better leverage the benefits and opportunities offered by socio-technical systems, work to address their risks, and restore human agency and accountability in better service of the public interest.

“

Employers aren’t aware of the information they need, there’s no duty to gather it, and in turn trade unions can’t see it.

Equality Task Force member

Endnotes

1 'The Privilege of Public Service' Ditchley Annual Lecture by the Chancellor of the Duchy of Lancaster, Michael Gove, 27th June 2020. https://www.gov.uk/government/speeches/the-privilege-of-public-service-given-as-the-ditchley-annual-lecture (accessed October 1, 2020).

2 The ETF's terms of reference are published online at https://www.ifow.org/news/2019/10/17/9pcsiw0vnn6ttia0ovtybsolz7svo1 (accessed October 1, 2020) and were to examine the following questions: What are the primary opportunities and challenges to equality arising from use of AI in the workspace? How do the provisions of Equality and Data Protection law apply to common uses of AI in the workspace? What is the role of individual rights in promoting equality through the age of AI? What are the roles of other duties and mechanisms to promote equality in the workspace? Are any new regulatory, guidance or other measures needed to achieve the purpose of the Equality Act in the workspace?

3 Please see the case studies, on which this analysis was founded. Simons, Josh, Graham Logan, Anna Thomas 'Machine Learning Case Studies' IFOW. https://www.ifow.org/publications/2020/2/24/machine-learning-case-studies (accessed October 1, 2020).

4 Gilbert, Abigail, Anna Thomas, Samuel Atwell, Joshua Simons (2020) 'The Impact of Automation on Labour Markets: Interactions with COVID19' IFOW. https://www.ifow.org/publications/2020/7/31/the-impact-of-automation-on-labour-markets-interactions-with-covid-19 (accessed October 1, 2020).

5 The reconvened expert group from the IFOW Commission met in June 2020. New Commissioners were Tabitha Goldstaub, Chair of the AI Council; Sir Michael Marmot, professor of Epidemiology; Kate Bell, Head of Economics Rights and TUC and Val Cooke, retail worker and USDAW representative.

6 Frank, Morgan R., David Autor, James E. Bessen, Erik Brynjolfsson, Manuel Cebrian, David J. Deming, Maryann Feldman et al. "Toward understanding the impact of artificial intelligence on labor." Proceedings of the National Academy of Sciences 116, no. 14 (2019): 6531-6539.

7 There is an emerging consensus, both nationally and internationally, that AI regulation is needed to enhance structures of accountability, and that this regulation must pay particular heed to equality.

8 Simons, Josh, Graham Logan, Anna Thomas 'Machine Learning Case Studies' IFOW. https://www.ifow.org/publications/2020/2/24/machine-learning-case-studies (accessed October 1, 2020).

9 Bathaee, Yavar. "The artificial intelligence black box and the failure of intent and causation." Harv. JL & Tech. 31 (2017): 889.

10 As defined in the Equality Task Force Terms of Reference, presented above and published at: https://www.ifow.org/news/2019/10/17/9pcsiw0vnn6ttia0ovtybsolz7svo1 (accessed October 1, 2020).

11 This has included critique of amendments and probing amendments on protecting equality as part of data protection law and on algorithmic, accountability fairness and impact assessments.

12 Josh Simons, Adrian Weller, Sophia Adams Bhatti 'Machine Learning and the Politics of Equal Treatment' Forthcoming, under review at FACT.

13 Today, 294 billion emails are sent every day, 4 petabytes of data are created on Facebook, and 5 billion searches are made. If you attempted to download all the data that is estimated to exist in the world much at the current average internet connection speed, it would take you about 1.8 billion years. The world's "global datasphere" (the total of all data created, captured or created) is expected to reach about 18 zettabytes of data by 2025 (a zettabyte is about 1,000 exabytes, which is 1,000 petabytes, which is 1,000 terabytes, which is 1,000 gigabytes). As the quantity and availability of data has proliferated almost exponentially, so the techniques of AI and machine learning have become more sophisticated. Bernard Marr, "How Much Data Is There In the World?," Blog, Bernard Marr, 2018. https://www.bernardmarr.com/default.asp?contentID=1846 (accessed October 1, 2020).

14 Such language can be found in a wide range of ML business promotional literature, and the CDEI have implied this: 'The use of algorithms has the potential to improve the quality of decision-making by increasing the speed and accuracy with which decisions are made. If designed well, they can reduce human bias in decision-making processes'. https://www.gov.uk/government/publications/interim-reports-from-the-centre-for-data-ethics-and-innovation/interim-report-review-into-bias-in-algorithmic-decision-making (accessed October 1, 2020).

15 Tom M. Mitchell, *Machine Learning*, New York, McGraw-Hill (1997).

16 Our research found that many of the most important practical applications of ML deploy supervised learning, in which a human specifies an outcome of interest – the thing an ML model learns to predict – and assembles the data from which the model learns to predict that outcome. Supervised ML is like telling an algorithm what it should learn to do and providing the examples from which it should learn to do it, except there are often thousands, millions, or even billions of examples.

Endnotes

17 It is important to note that this is not just about unrepresentative data. Unrepresentative data excludes some groups, perhaps because there aren't enough data points about that group, or the data points that exist are less numerous or rich. This is an important problem. But the problems of discrimination in AI would remain even in the most accurate data sets. Accurate data sets capture the structure of our social world more precisely, including the inequalities and injustices that characterize it. A representative dataset is not devoid of injustice. Cynthia Dwork and Deirdre Mulligan, "It's Not Privacy, and It's Not Fair," Stanford Law Review 66 (September 2013): 35–40; Kate Crawford, "The Hidden Biases in Big Data," Harvard Business Review, April 1, 2013, https://hbr.org/2013/04/the-hidden-biases-in-big-data (accessed October 1, 2020); David Garcia et al., "Analyzing Gender Inequality through Large-Scale Facebook Advertising Data," *Proceedings of the National Academy of Sciences of the United States of America* 115, no. 27 (2018): 6958–63.

18 Adams-Prassl, Jeremias 'Algorithmic management and the EU social acquis: opening the black box' A Thematic Working Paper for the Annual Conference of the European Centre of Expertise (ECE) in the field of labour law, employment and labour market policies: Exploring ways to improve working conditions of platform workers: The role of EU labour law' Session 2: Regulating and negotiating the 'blackbox' of algorithmic management. Directorate General for Employment, Social Affairs and Inclusion. March 2020.

19 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 (accessed October 1, 2020); Department of Housing and Urban Development, "Charge of Discrimination," 2019, https://www.hud.gov/press/press_releases_media_advisories/HUD_No_19_035 (accessed October 1, 2020).

20 Joshua Simons, "Equality in Machine Learning: Positive Duties and Indirect Discrimination in the Governance of Machine Learning," *SSRN Electronic Journal*, 2020;
Miranda Bogen, "All the Ways Hiring Algorithms Can Introduce Bias'," *Harvard Buisiness Review*, May 6, 2019, https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias?referral=03759&cm_vc=rr_item_page.bottom (accessed October 1, 2020);
Robin Allen and Dee Masters, "Artificial Intelligence: The Right to Protection from Discrimination Caused by Algorithms, Machine Learning and Automated Decision-Making" (EPA Forum, 2019), https://link.springer.com/article/10.1007/s12027-019-00582-w (accessed October 1, 2020);
Robin Allen, "Artificial Intelligence, Machine Learning, Algorithms and Discrimination Law: The New Frontier" (Discrimination Law in 2020, Congress House, 2020), https://482pe539799u3ynseg2hl1r3-wpengine.netdna-ssl.com/wp-content/uploads/2020/02/Discrimination-Law-in-2020.FINAL_-1.pdf (accessed October 1, 2020).

21 This distinction between technical bias and the reproduction of inequality is important. Bias is a well-defined word that involves distortions in how a machine learning model predicts its target variable – such as if the model is not well-calibrated or training data is insufficiently expressive. Bias is not the replication, reinforcement, and entrenching of historic patterns of social inequality. Bias should not be used as a euphemism for inequality.
Simons, "Equality in Machine Learning" forthcoming;
Sandra G. Mayson, "Bias In, Bias Out," *Yale Law Journal* 128, no. 8 (2019): 2218-;
Manish Raghavan et al., "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices," 2019.

22 Campolo, Alexander, and Kate Crawford. "Enchanted Determinism: Power without Responsibility in Artificial Intelligence." *Engaging Science, Technology, and Society* 6 (2020): 1-19.

23 HANA, an ML cloud powered database platform is used by Walmart to monitor their entire operations functioning in fine grained, real time detail. As outlined in Dyer-Witheford, Nick, Atle Mikkola Kjøsen, and James Steinhoff. "Inhuman power." *Artificial intelligence and the future of capitalism. London: Pluto Press* (2019).

24 Kellogg, Katherine C., Melissa A. Valentine, and Angele Christin. "Algorithms at work: The new contested terrain of control." *Academy of Management Annals* 14, no. 1 (2020): 366-410.

25 IFOW research into deployment of technology in Key Work Sectors, due to be published in early 2021 .

26 See: Wachter, Sandra. "Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR." *Computer law & security review* 34, no. 3 (2018): 436-449; Wachter, Sandra. "Affinity profiling and discrimination by association in online behavioural advertising." Berkeley Technology Law Journal 35, no. 2 (2020); Ducato, Rossana, Miriam Kullmann, and Marco Rocca. "Customer ratings as a vector for discrimination in employment relations? Pathways and pitfalls for legal remedies." In *Proceedings of the Marco Biagi Conference*. 2018.

27 Simons, Thomas, and Graham, "Machine Learning at Work: Case Studies";
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 (accessed October 1, 2020);
Department of Housing and Urban Development, "Charge of Discrimination," 2019, https://www.hud.gov/press/press_releases_media_advisories/HUD_No_19_035 (accessed October 1, 2020).

Endnotes

28 The Royal Society 'Explainable AI: The Basics' (2019)
https://royalsociety.org/topics-policy/projects/explainable-ai/?gclid=Cj0KCQjwtsv7BRCmARIsANu- CQdQo8YqHPkE0NlV_0OojOxvCsLiMy1W3Z1Z86CDcKNyhgOD-XIpqKEaArLxEALw (accessed October 1, 2020).

29 Simons, "Equality in Machine Learning"; Mayson, "Bias In, Bias Out"; Ruha Benjamin, Race after Technology: Abolitionist Tools for the New Jim Code (Medford, MA: Polity, 2019); Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," California Law Review 104, no. 3 (June 1, 2016): 671–732; Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (New York: Crown, 2016).

30 Graham, Logan, Abigail Gilbert, Joshua Simons, Anna Thomas,
'Artificial Intelligence in Hiring: Assessing Impacts on Equality' (2020) IFOW.
https://www.ifow.org/publications/2020/4/27/artificial-intelligence-in-hiring-assessing-impacts-on-equality
(accessed October 1, 2020).

31 Crawford, "The Hidden Biases in Big Data"; Mayson, "Bias In, Bias Out."

32 People often need to assemble and label the data from which a model learns, building in all kinds of important assumptions. To build a classifier that identifies 'fake' C.V.s submitted into a large employer, thousands of CVs would have to be labelled as 'fake' and 'not fake'. Social media companies use a labelling process to build datasets to train machine learning models to predict whether content is false or misleading.
Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," Jenna Burrell, 2016, http://journals.sagepub.com/doi/abs/10.1177/2053951715622512 (accessed October 1, 2020);
Kiel Brennan-Marquez, "'Plausible Cause': Explanatory Standards in the Age of Powerful Machines,"
Vanderbilt Law Review 70, no. 4 (2017): 1249–1301.

33 We return to this point since a great deal of the debate about discrimination in ML has focused on the inclusion and exclusion of features, to the detriment of more significant issues. Jon Kleinberg and Sendhil Mullainathan, "Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability," September 12, 2018. https://arxiv.org/abs/1809.04578 (accessed October 1, 2020).

34 Alexandra Chouldechova et al., "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions," *Proceedings of Machine Learning Research* 81 (2018): 134–48;
Rhema Vaithianathan et al., "Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation" (Allgheny County, Pennsylvania: Allgheny County Analytics, April 2019), https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf (accessed October 1, 2020).

35 Graham, Logan, Abigail Gilbert, Joshua Simons, Anna Thomas,
'Artificial Intelligence in Hiring: Assessing Impacts on Equality' (2020) IFOW.
https://www.ifow.org/publications/2020/4/27/artificial-intelligence-in-hiring-assessing-impacts-on-equality
(accessed October 1, 2020).

36 Ibid 35.

37 In a 2018 survey by LinkedIn, 68% of businesses reported that they were using AI to save them time. In the same survey, 78% said they were tackling hiring 'diverse talent' head on either to improve culture (78%) or to boost financial performance (62%). 43% of those surveyed thought that using AI would help them 'remove human bias'.

38 McNamara, A., Smith, J., Murphy-Hill, E. (2018).
Does ACM's code of ethics change ethical decision making in software development?" In G. T. Leavens, A. Garcia, C. S. Păsăreanu (Eds.) Proceedings of the 2018 26th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering—ESEC/FSE 2018 (pp. 1–7). New York: ACM Press. In brief, their finding was that the effectiveness of guidelines or ethical codes is almost zero and that they do not change the behaviour of professionals from the tech community. Cited in Hagendorff, Thilo. "The ethics of Ai ethics: An evaluation of guidelines." *Minds and Machines* (2020): 1-22.

39 These will be further complicated by intellectual property rights which may exist, subject to limitations, in relation to information which has an economic value which requires economic investments. We note however that the nature of personal data reduces the absolute nature of IPRs and our recommendations take this into account. Banterle, Francesco. "The interface between data protection and IP law: the case of trade secrets and the database sui generis right in marketing operations, and the ownership of raw data in big data analysis." In Personal Data in Competition, Consumer Protection and Intellectual Property Law, pp. 411-443. Springer, Berlin, Heidelberg, 2018.

40 See for instance Green, Ben. "Data science as political action: grounding data science in a politics of justice." Available at SSRN 3658431 (2020).

41 Houghton, Ed and Louisa Houghton 'Workplace Technology: The Employee Experience' CIPD (2020).
https://www.cipd.co.uk/Images/workplace-technology-1_tcm18-80853.pdf (accessed October 1, 2020).

Endnotes

42 Levy, Karen, and Solon Barocas. "Privacy at the Margins| refractive surveillance: Monitoring customers to manage workers." *International Journal of Communication* 12 (2018): 23; Kellogg, Katherine C., Melissa A. Valentine, and Angele Christin. "Algorithms at work: The new contested terrain of control." *Academy of Management Annals* 14, no. 1 (2020): 366-410.

43 See case by Uber Drivers to make algorithm transparent, https://www.theguardian.com/technology/2020/jul/20/uber-drivers-to-launch-legal-bid-to-uncover-apps-algorithm (accessed October 1, 2020);
Prospect Union Survey, (https://prospect.org.uk/about/future-of-work-technology-and-data/ (accessed October 1, 2020);
forthcoming IFOW surveys of retail workers; Jennifer Cobbe, Chris Norval, and Jatinder Singh, "What Lies beneath: Transparency in Online Service Supply Chains," *Journal of Cyber Policy: Special Issue: Consolidation of the Internet* 5, no. 1 (2020): 65–93;
Jennifer Cobbe, "Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making," *Legal Studies (Society of Legal Scholars)* 39, no. 4 (2019): 636–655, https://doi.org/10.1017/lst.2019.9 (accessed October 1, 2020);
Cobbe; Danielle Citron and Frank Pasquale, "The Scored Society: Due Process for Automated Predictions," *Washington Law Review* 89, no. 1 (2014): 1–33.

44 Custers, Bart, and Helena Ursic. "Worker Privacy in a Digitalized World Under European Law." Comp. Lab. L. & Pol'y J. 39 (2017): 323; Wachter, Sandra. "Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR." Computer law & security review 34, no. 3 (2018): 436-449;
Levy, Karen, and Solon Barocas. "Privacy at the Margins| refractive surveillance: Monitoring customers to manage workers." International Journal of Communication 12 (2018): 23.

45 Survey conducted by USDAW union of their members, in field August–November 2020. Total N=1015, still in field. Developed with the support of IFOW for work funded by Trust for London.

46 11% of UK workers have earned income from working on digital labor platforms (Huws, U., Spencer, N., & Joyce, S. (2016). Crowd work in Europe: Preliminary results from a survey in the UK, Sweden, Germany, Austria and the Netherlands.) and it is predicted that a third of jobs will be mediated by online platforms by 2025 (Standing, 2016 cited in Graham, Mark, and Jamie Woodcock. "Towards a fairer platform economy: introducing the Fairwork Foundation." Alternate Routes 29 (2018).

47 Ellen Sheng. Employee Privacy in The U.S. is at Stake as Corporate Surveillance Technology Monitors Workers' Every Move, CNBC (Apr. 15, 2019), cited in Nelson, Josephine, Management Culture and Surveillance (December 16, 2019). 43 Seattle U. L. Rev. 2, 631 (2020).

48 Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. "Mitigating bias in algorithmic hiring: Evaluating claims and practices." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 469-481. 2020.

49 Andrew D. Selbst and Solon Barocas, "The Intuitive Appeal of Explainable Machines," *SSRN Electronic Journal*, 2018, https://doi.org/10.2139/ssrn.3126971 (accessed October 1, 2020);
Edwards, Lilian, and Michael Veale. "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for." Duke L. & Tech. Rev. 16 (2017): 18.;
Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law* 7, no. 2 (2017): 76–99.

50 The Royal Society 'Explainable AI: The Basics' (2019). https://royalsociety.org/topics-policy/projects/explainable-ai/?gclid=Cj0KCQjwtsv7BRCmARIsANu-CQdQo8YqHPkE0NlV_0OojOxvCsLiMy1W3Z1Z86CDcKNyhgOD-XIpqKEaArLxEALw_wcB (accessed October 1, 2020).

51 Singh, Jatinder, Jennifer Cobbe, and Chris Norval. "Decision provenance: Harnessing data flow for accountable systems." IEEE Access 7 (2018): 6562-6574.

52 Dencik, Lina, Arne Hintz, Joanna Redden, and Emiliano Treré. "Exploring data justice: Conceptions, applications and directions." (2019): 873-881.

53 Kaminski, Margot E., and Gianclaudio Malgieri. "Multi-layered explanations from algorithmic impact assessments in the GDPR." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 68–79. 2020;
Kaminski, Margot E., and Gianclaudio Malgieri. "Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations." *Available at SSRN 3456224* (2019).

54 Lynskey, Orla. "Deconstructing data protection: the 'added-value'of a right to data protection in the EU legal order." International & Comparative Law Quarterly 63.3 (2014): 569–597.

55 Ibid 53.

56 See articles 13/14.

Endnotes

57 We note criticism includes a lack of public awareness and understanding of their rights, (see for instance Sideri, Maria, and Stefanos Gritzalis. "Are We Really Informed on the Rights GDPR Guarantees?" In International Symposium on Human Aspects of Information Security and Assurance, pp. 315-326. Springer, Cham, 2020); and business approaches to DPIAs and their completion; (see for instance Kröger, Jacob Leon, Jens Lindemann, and Dominik Herrmann. "How do app vendors respond to subject access requests? A longitudinal privacy study on iOS and Android Apps." In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, pp. 1–10. 2020).

58 Algorithm Watch have launched a global inventory of the various frameworks that seek to set out principles of how systems for automated decision making (ADM) can be developed and implemented ethically. https://inventory.algorithmwatch.org/ (accessed October 1, 2020).

59 Institute of Electrical and Electronics Engineers. (2018). The IEEE Global Initiative on ethics of autonomous and intelligent systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf (accessed October 1, 2020); High Level Expert Group on AI; European Commission. (2019). Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed October 1, 2020); Floridi, L., & Lord Clement-Jones. (2019, March 20). The five principles key to any ethical framework for AI. New Statesman. https://tech.newstatesman.com/policy/ai-ethics-framework (accessed October 1, 2020); Leslie, D. (2019a). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/ZENODO.3240529 (accessed October 1, 2020); Organisation for Economic Co-operation and Development. (2019a). OECD principles on AI. https://www.oecd.org/going-digital/ai/principles/ (accessed October 1, 2020).

60 Whereas the standards, certification, and governance of digital and software-driven innovation have previously concentrated on the technical dynamics and specifications necessary to assure the performance and safety of these sorts of technologies, new concerns about the social shaping and ethical consequences of the processes and products of AI research and innovation have now come to the forefront, as seen in recent ICO Guidance.

61 IEEE Standards Project for Transparency of Autonomous Systems provides a Standard for developing autonomous technologies that can assess their own actions and help users understand why a technology makes certain decisions in different situations. See more at: https://ethicsinaction.ieee.org/p7000/ (accessed October 1, 2020).

62 In collaboration with the International Electrotechnical Commission (IEC).

63 ISO/IEC JTC 1 SC42/WG3.

64 Information Commissioner's Office and The Alan Turing Institute, "Explaining Decisions Made with AI," 2020, https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/ (accessed October 1, 2020).

65 Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute. https://doi.org/10.5281/ZENODO.3240529 (accessed October 1, 2020).

66 The Centre for Data Ethics and Innovation's Approach to the Governance of Data Driven Technology, 19th July 2019. https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovations-approach-to-the-governance-of-data-driven-technology/the-centre-for-data-ethics-and-innovations-approach-to-the-governance-of-data-driven-technology (accessed October 1, 2020).

67 Ellen Sheng. Employee Privacy in The U.S. is at Stake as Corporate Surveillance Technology Monitors Workers' Every Move, CNBC (Apr. 15, 2019), cited in Nelson, Josephine, Management Culture and Surveillance (December 16, 2019). 43 Seattle U. L. Rev. 2, 631 (2020).

68 For more insight into the miriad of modes of AI governance and their relative success, see: Koene, Ansgar, Chris Clifton, Yohko Hatada, Helena Webb, and Rashida Richardson. "A governance framework for algorithmic accountability and transparency." (2019).

69 Hagendorff, Thilo. "The ethics of Ai ethics: An evaluation of guidelines." *Minds and Machines* (2020): 1-22.

70 As we documented in our interim report on hiring tools, Graham, Logan, Abigail Gilbert, Joshua Simons, Anna Thomas 'Artificial Intelligence in Hiring: Assessing Impacts on Equality' (2020). https://www.ifow.org/publications/2020/4/27/artificial-intelligence-in-hiring-assessing-impacts-on-equality (accessed October 1, 2020).

71 Tisne, Martin. 'The Data Delusion: Protecting Individual Data Isn't Enough When the Harm is Collective' (2020). https://cyber.fsi.stanford.edu/publication/data-delusion (accessed October 1, 2020).

72 We recommend an Equality Impact Assessment as a voluntary interim measure, see here: https://www.ifow.org/consultation (accessed October 1, 2020).

Endnotes

73 This is informed by IFOW workshops with the Financial Times and IET.

74 Yvette D. Clarke, Cory Booker, and Ron Wyden, "Algorithmic Accountability Act of 2019," Pub. L. No. H.R.2231 (2019), https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202019%20Bill%20Text.pdf (accessed October 1, 2020);
Justice B.N. Srikrishna, "A Free and Fair Digital Economy: Protecting Privacy, Empowering Indians" (Committee of Experts under the Chairmanship of Justice B.N. Shrikrishna, July 2018), https://meity.gov.in/writereaddata/files/Data_Protection_Committee_Report.pdf (accessed October 1, 2020);
"Creating a French Framework to Make Social Media Companies More Accountable: Interim Mission Report" (France: The French Secratary of State for Digital Affairs, 2019), http://www.iicom.org/images/iic/themes/news/Reports/French-social-media-framework---May-2019.pdf (accessed October 1, 2020);
"The Digital Services Act," Pub. L. No. COM (2020) 37 (2020).

75 The kind of accountability legislation proposed in this report that would be pursued under a Biden administration could impact that proposed in this report. The legislation proposed in both jurisdictions could be developed together.

76 Consider Thor's job application screening system from our machine learning case studies. This system raises data protection questions about how data from the social media profiles of existing employees was acquired, and whether consent could reasonably be given in the context of the hierarchical relationship between employer and employee. The system also raises equality law questions about whether past data about customer satisfaction is a reasonable information base on which to train the system's machine learning model, and whether future customer satisfaction is a defensible outcome to predict and use to sort which applicants will be invited for interview.

77 The Acas Code of Practice on Disclosure of Information to Trade Unions for Collective Bargaining Purposes came into effect on 22 August 1977 and was issued under section 6 of the Employment Protection Act 1975 (now section 199 of the Trade Union and Labour Relations (Consolidation) Act 1992 ("the 1992 Act"). Revised code available online at: https://archive.acas.org.uk/media/273/Code-of-Practice---Disclosure-of-information-to-trade-unions/pdf/11287_CoP2_Collective_Bargaining_v1_0_Accessible.pdf (accessed October 1, 2020).

78 In the Competition Act 1998 and Enterprise Act of 2002, as enforced by the Competition and Markets Authority (CMA).

79 Todolí-Signes, Adrián. "Algorithms, artificial intelligence and automated decisions concerning workers and the risks of discrimination: the necessary collective governance of data protection."
Transfer: European Review of Labour and Research 25, no. 4 (2019): 465-481.

80 Some courts regard inferential data as 'second grade' personal data, and may accord it slightly less protection (Wachter, Sandra, and Brent Mittelstadt. "A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI." Colum. Bus. L. Rev. (2019): 494).

81 Veale, Michael, Reuben Binns, and Lilian Edwards.
"Algorithms that remember: model inversion attacks and data protection law." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (2018): 20180083.

82 GDPR's Article 80 (2) granted Member States the choice to allow class actions that claim monetary damages. Some countries, such as France, Austria, and Belgium decided to allow damage claims (Alexia Pato, The national adaptation of Article 80 GDPR: towards the effective private enforcement of collective data protection rights, in National adaptations of the GDPR, p.106).

83 Data Protection Act (2018). https://www.gov.uk/data-protection (accessed October 1, 2020).

84 Adams-Prassl, Jeremias 'Algorithmic management and the EU social acquis: opening the black box' A Thematic Working Paper for the Annual Conference of the European Centre of Expertise (ECE) in the field of labour law, employment and labour market policies: Exploring ways to improve working conditions of platform workers: The role of EU labour law' Session 2: Regulating and negotiating the 'blackbox' of algorithmic management. Directorate General for Employment, Social Affairs and Inclusion. March 2020.

85 Article 22 of GDPR 'Automated Decision Making, Including Profiling'. https://gdpr-info.eu/art-22-gdpr/ (accessed October 1, 2020).

86 ICO Guidance, available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-does-the-gdpr-say-about-automated-decision-making-and-profiling/ (accessed October 1, 2020).

87 Recital 71 explicitly lists 'e-recruiting practices without any human intervention' cited in Prassl, Jeremias 'Algorithmic management and the EU social acquis: opening the black box' A Thematic Working Paper for the Annual Conference of the European Centre of Expertise (ECE) in the field of labour law, employment and labour market policies: Exploring ways to improve working conditions of platform workers: The role of EU labour law' Session 2: Regulating and negotiating the 'blackbox' of algorithmic management. Directorate General for Employment, Social Affairs and Inclusion. March 2020.

Endnotes

88 Recital 85 of the GDPR. See ICO Guidance 'Guide to the General Data Protection Regulation (GDPR) on this at: https://ico.org.uk/media/for-organisations/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf (accessed October 1, 2020).

89 GDPR as set out in Article 22(2) a-c.

90 For the way challenges around consent are compounded by the Industrial internet of things, see e.g. Wachter, Sandra. "Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR." *Computer law & security review* 34, no. 3 (2018): 436-449.

91 The 'human in command' and consultative approach, was initially advocated by the European Economic and Social Committee (see Opinion of the European Economic and Social Committee on 'Artificial intelligence — The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society' (own-initiative opinion) (2017/C 288/01) Rapporteur: Catelijne MULLER) and has been endorsed by the ILO (see 'Negotiating the algorithm: Automation, artificial intelligence and labour protection' EMPLOYMENT Working Paper No. 246, ILO) and UNI Global Union (Union, UNI Global. "Top 10 Principles for Ethical Artificial Intelligence." *Nyon, Switzerland* (2017).

92 The ICO maintains a list of examples of processing likely to result in high risk: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias/examples-of-processing-likely-to-result-in-high-risk/ (accessed October 1, 2020).

93 The 'high risk' threshold covers each key decision in our case studies, although this is not widely understood. To assess whether something is 'high risk', the GDPR is clear that you need to consider both the likelihood and severity of any potential harm to individuals. 'Risk' implies a more than remote chance of some harm. 'High risk' implies a higher threshold, either because the harm is more likely, or because the potential harm is more severe, or a combination of the two. Assessing the likelihood of risk in that sense is part of the job of a DPIA.

94 Binns, Reuben. "Data protection impact assessments: a meta-regulatory approach." International Data Privacy Law 7, no. 1 (2017): 22-35.

95 Insight from Task Force members.

96 This includes the vague 'meaningful information about the logic involved' which we anticipate will be tested in the courts. Edwards, Lilian, and Michael Veale. "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for." Duke L. & Tech. Rev. 16 (2017): 18.

97 Mahieu, René, and Jef Ausloos. "Recognising and Enabling the Collective Dimension of the GDPR and the Right of Access." (2020).

98 Edwards, Lilian, and Michael Veale. "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for." Duke L. & Tech. Rev. 16 (2017): 18; Veale, Michael, Reuben Binns, and Lilian Edwards. "Algorithms that remember: model inversion attacks and data protection law." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376, no. 2133 (2018): 20180083.

99 Kröger, Jacob Leon, Jens Lindemann, and Dominik Herrmann. "How do app vendors respond to subject access requests? A longitudinal privacy study on iOS and Android Apps." In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, pp. 1–10. 2020.

100 Kröger, Jacob Leon, Jens Lindemann, and Dominik Herrmann. "How do app vendors respond to subject access requests? A longitudinal privacy study on iOS and Android Apps." In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, pp. 1–10. 2020.

101 On the basis of Union or Member State law, which in the case of the UK is specified under Schedule 1 of the DPA 2018.

102 The Equality Act (2010) Explanatory Notes. http://www.legislation.gov.uk/ukpga/2010/15/notes/contents (accessed October 1, 2020).

103 Thomas Giegerich (Editor) The European Union as Protector and Promoter of Equality. Springer, Switzerland.

104 The UK is different in this important respect from the U.S., which in many instances does require the demonstration of the subjective mental state of intent as a necessary condition for establishing direct discrimination. McDonnell Douglas Corp v Green 411 US 792 (1973) 802–3. More recent attempts to adjust evidential rules to assume discriminatory intent if certain objective criteria are satisfied have largely been reversed by U.S. courts. St Mary's Honor Center v Hicks 509 US 502 (1993) 519.

105 JFS Case; Lord Nicholls has described a broad notion of subjective intent, which includes subconscious intent. This raises questions about the plausible meaning of intent that may become especially salient when analysing the choices of an organisation about the design and deployment of a machine learning model. Constable of West Yorkshire Police v Khan [2001] UKHL 48. Sheila Foster, "Causation in Antidiscrimination Law: Beyond Intent versus Impact," *Houston Law Review* 41, no. 5 (2005): 1469–1548; David Strauss, "Discriminatory Intent and the Taming of Brown," *University of Chicago Law Review* 56, no. 3 (1989): 935–935.

Endnotes

106 A useful gloss on the statutory test was outlined by Lord Goff in James v Eastleigh Borough Council [1990] 2 AC 751, 774. It asks: "Would the complainant have received the same treatment … but for his or her sex?" The modern approach to the need for discriminatory intent was discussed in Onu v Akwiwu [2014] 1 WLR 3636. The decision in R (on the application of E) v Governing Body of JFS [2009] UKSC 15 confirmed that courts will impose this objective test, in part because of difficulties in the standard of proof required to demonstrate discriminatory intent.

107 Reva B. Siegel, "Blind Justice: Why The Court Refused to Accept Statistical Evidence of Discriminatory Purpose in McCleskey v. Kemp - and Some Pathways for Change," *Northwestern University Law Review* 112, no. 6 (2018): 1269–1291; Barbara D. Underwood, "Law and the Crystal Ball: Predicting Behavior with Statistical Inference and Individualized Judgment," *The Yale Law Journal* 88, no. 7 (1979): 1408–1448; Whereas in *Coll* the factors that caused the disadvantage to women were in the defendant's control. Lady Hale, R (on the application of Coll) v Secretary of State for Justice (UKSC May 24, 2017). Lady Hale: "… the question of comparing like with like must always be treated with great care – men and women are different from one another in many ways, but that does not mean that the relevant circumstances cannot be the same for the purpose of deciding whether one has been treated less favourably than the other. Usually, those circumstances will be something other than the personal characteristics of the men and women concerned, something extrinsic rather than intrinsic to them. Simons, "Equality in Machine Learning" forthcoming;
Allen, "Artificial Intelligence, Machine Learning, Algorithms and Discrimination Law: The New Frontier";
Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI," SSRN, 2020; B. A. *Hepple, Equality: The Legal Framework* (Oxford, United Kingdom: Hart Publishing, 2014).

108 R v JFS

109 Throughout the *Declaration of Principles on Equality*, the concept of equality, as well as its equivalent, "full and effective equality", has a content which is larger than that of "non-discrimination".

110 For instance, it is doubtful that the logic of Coll applies to machine learning cases such as Networkz. The most salient consideration is the "exact correspondence" test. Not *all* women suffer the relative disadvantage of being shown lower-paid job adverts in Networkz's case – in fact, those who have tended to click on higher paid job adverts than men may be shown job adverts with an average income that is higher than the average income of the job ads shown to men as a whole. Because the machine learning system is personalized, the disadvantage is centered on averages: women *on average*, are shown job adverts with lower average incomes than those shown to men. Machine learning systems will rarely meet the "exact correspondence" test. Dee Masters, "Identifying Direct Discrimination in 'Proxy Cases' after R (on the Application of Coll) v Secretary of State for Justice," Cloisters – Barristers Chambers, May 31, 2017, https://www.cloisters.com/identifying-direct-discrimination-in-proxy-cases-after-r-on-the-application-of-coll-v-secretary-of-state-for-justice/ (accessed October 1, 2020).

111 As Lady Hale explains: "It is commonplace for the disparate impact, or particular disadvantage, to be established on the basis of statistical evidence." Lady Hale, Essop v Home Office (UKSC 2017).

112 In Seymour-Smith, the ECJ adopted two distinct standards: if "the statistics…indicate that a considerably smaller percentage" of the protected than the comparator group satisfy the relevant requirement; or if "the statistical evidence reveal[s] a lesser but persistent and relatively constant disparity over a long period" between the proportion of the two groups who satisfy the relevant requirement. R v Secretary of State for Employment, ex parte Seymour-Smith and Perez, C-167/97 253 (I.R.L.R. 1999). The Supreme Court (per Hale LJ) has described the distinct purposes of prohibitions against direct and indirect discrimination in this way (at 57).

113 Our case studies draw attention to the need to identify and consider what the underlying inequalities are, and what causes them, as part of the process of evaluation. For example, the reason Thor's job tenure prediction model consistently ranks women below men is that the training data records that, on average, women have stayed in data science and engineering jobs less long than men. There are likely to be multiple reasons for this disparity, including Thor's own policies about paid maternity and paternity leave and government policies concerning childcare and early years education. Thor might argue that the underlying cause is outside their control. An employee might argue that job tenure is not a reasonable proxy for the future performance of employees and that even if it is, past data about job tenure is not a reasonable basis for predicting the job tenure of future employees.

114 Bilka-Kaufhaus GmbH v Weber von Hartz [1986] IRLR 317.

115 Bank Mellat. In the UK, the intensity of judicial scrutiny of the justification defence is sensitive to several factors, including: the nature of the duty infringed, the protected ground in question, the nature of the infringer (in particular: whether it is a public or private person and the degree to which it is able to bear the cost of the actions required to achieve non-discrimination) Khaitan, T. (2015-05-01). The Architecture of Discrimination Law. In (Ed.), A Theory of Discrimination Law: Oxford University Press,. From http://www.oxfordscholarship.com.ezp-prod1.hul.harvard.edu/view/10.1093/acprof:oso/9780199656967.001.0001/acprof-9780199656967-chapter-3 (accessed October 1, 2020).

116 Ken, Robin and Dee Masters, "Artificial Intelligence: The Right to Protection from Discrimination Caused by Algorithms, Machine Learning and Automated Decision-Making" 595.

117 Seldon v Clarkson Wright and Jakes [2012] UKSC 16, [2012] ICR 716.

Endnotes

118 Cobbe, Jennifer, Chris Norval, and Jatinder Singh. "What lies beneath: transparency in online service supply chains." Journal of Cyber Policy 5, no. 1 (2020): 65-93.

119 Graham, Logan, Anna Thomas, Joshua Simons, Abigail Gilbert .
'Artificial Intelligence in Hiring: Assessing Impacts on Equality' IFOW.
https://www.ifow.org/publications/2020/4/27/artificial-intelligence-in-hiring-assessing-impacts-on-equality (accessed October 1, 2020).

120 A particular problem re ML is that is attempts to 'automate' the process of discovering relationships in a way that makes qualitative judgments about the relationship hard and sometimes impossible.

121 Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. "Mitigating bias in algorithmic hiring: Evaluating claims and practices." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 469-481. 2020. https://arxiv.org/pdf/1906.09208.pdf (accessed October 1, 2020).

122 Similarly, an ML system may aid the search for a 'less discriminatory' means to achieve the same purpose but cannot purport to determine what that is without a human. The search, and evaluation, must be qualitative as well as quantitive, involving objective and subjective elements.

123 Schedule 8, Para 20, Equality Act 2010.

124 Part 5 of the Code. Note The Act imposes some restrictions on the types of disability-related enquiries that can be made prior to making a job offer, although questions are generally permitted to determine whether reasonable adjustments need to be made in relation to an assessment.

125 In summary the PSED requires public bodies in performance of their functions, to have due regard to the need to:
- Eliminate unlawful discrimination, harassment and victimisation and other conduct prohibited by the Act.
- Advance equality of opportunity between people who share a protected characteristic and those who do not.
- Foster good relations between people who share a protected characteristic and those who do not.

126 Commissioners, and see for instance: Manfredi, Simonetta, Lucy Vickers, and Kate Clayton-Hathway.
"The public sector equality duty: enforcing equality rights through second-generation regulation."
Industrial Law Journal 47, no. 3 (2018): 365-398.

127 The Networkz case shows that whatever patterns of social inequality exist in the data on which a statistical model is trained will be reflected in its predictions. If those patterns of social inequality tend to correlate with protected characteristics like race and gender, which they tend to, as that's why those characteristics are protected in the first place, then the predictions of that statistical model will have disparate impact across protected groups. The model has merely learned that our society is ridden with systemic inequalities which shape the distribution of the outcome it has been asked to predict, in Network's case, click probability.

128 For instance, the South Africa Constitution was rewritten to include positive, remedial action. See more in Jadwanth, Saras. "Affirmative action in a transformative context: The South African experience." *Conn. L. Rev.* 36 (2003): 725.

129 Quote Hale LJ, as she cites Birmingham and EOC. The leading case on direct discrimination is *R v Birmingham City Council, ex p Equal Opportunities Commission* [1989] 1 AC 1155.

130 For instance, suppose Networkz were to impose a maximum disparity of 5 percent between the average incomes of job ads shown to men and women. This would involve making determinations because of a protected characteristic, distributing adverts on the basis of gender, potentially violating direct discrimination. We note that further work is required to determine the remit of exceptions for positive discrimination in the EA.

131 Graham, Logan, Anna Thomas, Joshua Simons, Abigail Gilbert.
'Artificial Intelligence in Hiring: Assessing Impacts on Equality' IFOW.
https://www.ifow.org/publications/2020/4/27/artificial-intelligence-in-hiring-assessing-impacts-on-equality (accessed October 1, 2020).

132 Owen M. Fiss, "Groups and the Equal Protection Clause," *Philosophy & Public Affairs* 5, no. 2 (1976): 107–177;
Iris Marion Young, *Justice and the Politics of Difference* (Princeton, N.J.: Princeton University Press, 1990); Iris Marion Young, "Equality of Whom? Social Groups and Judgments of Injustice," *The Journal of Political Philosophy* 9, no. 1 (2001): 1–18.

133 Mahieu, René, and Jef Ausloos. "Recognising and Enabling the Collective Dimension of the GDPR and the Right of Access." (2020);
Mahieu, René L P, Hadi Asghari, and Michel van Eeten. "Collectively Exercising the Right of Access: Individual Effort, Societal Effect." Internet Policy Review 7, no. 3 (2018): Available at: https://doi.org/10.14763/2018.3.927 (accessed October 1, 2020).

134 Detailed as Trend 5 in Thomas, Anna, Abigail Gilbert, Samuel Atwell, Joshua Simons.
'A better future for work: the world after Covid-19' Available at: https://www.ifow.org/publications/2020/6/10/a-rapid-review-with-the-future-of-work-commission-a-better-future-for-work-the-world-after-covid-19 (accessed October 1, 2020).

Endnotes

135 Graham, Logan, Anna Thomas, Joshua Simons, Abigail Gilbert.
'Artificial Intelligence in Hiring: Assessing Impacts on Equality' IFOW.
https://www.ifow.org/publications/2020/4/27/artificial-intelligence-in-hiring-assessing-impacts-on-equality
(accessed October 1, 2020).

136 Zuiderveen Borgesius, Frederik. 'Discrimination, artificial intelligence and algorithmic decision making'
Council of Europe (2020).
https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 (accessed October 1, 2020).

137 Ibid 136.

138 Ibid 136.

139 Kosta, E (2017) Surveilling Masses and Unveiling Human Rights – Uneasy Choices for the Strasborg Court.
Inaugural Address (Tilburg Law School Research Paper).

140 Analysis of case UK and ECJ caselaw use of the word and concept of 'fairness' in discrimination cases, compared to the technical metrics of fairness developed by the computer science community suggests that the courts embrace 'contextual equality' which relies on common knowledge and intuition. (Wachter, Sandra, Brent Mittelstadt, and Chris Russell.
"Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI." Available at SSRN (2020).
The same study highlighted that that some groups, especially intersectional groups, are especially vulnerable. Our review suggests that these groups also have the least protection and recourse to a remedy.

141 S. Fredman, "Addressing Disparate Impact: Indirect Discrimination and the Public Sector Equality Duty,"
Industrial Law Journal (London) 43, no. 3 (2014): 349–363.; 70
As we documented in our interim report on hiring tools, Graham, Logan, Abigail Gilbert, Joshua Simons,
Anna Thomas 'Artificial Intelligence in Hiring: Assessing Impacts on Equality' (2020).
https://www.ifow.org/publications/2020/4/27/artificial-intelligence-in-hiring-assessing-impacts-on-equality
(accessed October 1, 2020).

142 Additionally, there is no prohibition on the use of solely automated systems in the EA, to mirror the DPA. While the task force will explore collaborations which may be able to apply the levers in the DPA to address this, principally further legal infrastructure is required.

143 Equality Act 2020 Part 3, 29 (1) Provision of Services.

144 Graham, Logan, Anna Thomas, Joshua Simons, Abigail Gilbert.
'Artificial Intelligence in Hiring: Assessing Impacts on Equality' IFOW.
https://www.ifow.org/publications/2020/4/27/artificial-intelligence-in-hiring-assessing-impacts-on-equality
(accessed October 1, 2020).

145 As noted in part 3, there is no requirement in the EA for the private sector to actively promote equality of opportunity, or reduce inequalities of outcome. In this context incentives to monitor and track the evolving outcomes of ML tools in work are limited to corporate social responsibility choices.

146 Harris, Swee Leng. "Data Protection Impact Assessments as rule of law governance mechanisms." *Data & Policy* 2 (2020).

147 Graham, Logan, Anna Thomas, Joshua Simons, Abigail Gilbert.
'Artificial Intelligence in Hiring: Assessing Impacts on Equality' IFOW.
https://www.ifow.org/publications/2020/4/27/artificial-intelligence-in-hiring-assessing-impacts-on-equality
(accessed October 1, 2020).

148 'European Commission White Paper on Artificial Intelligence: Our Response' Scottish Government (August 2020)
https://www.gov.scot/publications/scottish-government-response-european-commission-white-paper-artificial-intelligence/
(accessed October 1, 2020).

149 See our consultation online at https://www.ifow.org/consultation and proposed EQIA at https://static1.squarespace.com/static/5aa269bbd274cb0df1e696c8/t/5ecfc7985080b661289f1cf5/1590675365747/EIA+template.

150 Response To The European Commission White Paper "On Artificial Intelligence – A European Approach To Excellence And Trust"
Robin Allen QC and Dee Masters, barristers, AI Law Consultancy (www.ai-lawhub.com) and Cloisters chambers
(www.cloisters.com).

151 Digital Regulation Cooperation Forum, launched July 2020.
https://www.gov.uk/government/publications/digital-regulation-cooperation-forum (accessed October 1, 2020).

152 Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529 (accessed October 1, 2020);
Recommendation of the Council on Artificial Intelligence OECD/LEGAL/0449
https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed October 1, 2020).

Endnotes

153 Around the world, governments have begun to develop legislative frameworks to structure accountability in the design and deployment of AI, ML, and algorithms. For instance, see: Clarke, Booker, and Wyden, Algorithmic Accountability Act of 2019; The Digital Services Act; "The Personal Data Protection Bill" (2018), https://meity.gov.in/writereaddata/files/Personal_Data_Protection_Bill,2018.pdf (accessed October 1, 2020); NITI Aayog, "National Strategy for Artificial Intelligence," Discussion Paper, June 2018, http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf (accessed October 1, 2020); "Creating a French Framework to Make Social Media Companies More Accountable: Interim Mission Report."

154 US Judiciary Committee report argued, the underlying purpose of competition law and anti-trust is also to hold corporations accountable, and to structure market power, in the public interest.

155 'Quality Assurance: Our Principles and Approach' Financial Ombudsman Service. https://www.financial-ombudsman.org.uk/files/3094/our-quality-assurance-principles.pdf (accessed October 1, 2020).

156 We propose using the concept and language of reasonable adjustments already in the Equality Act, but this needs further consultation and it is acknowledged that there is scope for improvement.

157 Binns, Reuben. "On the apparent conflict between individual and group fairness." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 514-524. (2020).

158 Technology can and should be harnessed to ensure efficiency of such activities. Precedents exist for multi-party API ecosystems (such as open banking) for businesses sharing data with certified bodies securely. Similar approaches could be adopted for creating comprehensive, responsive and lower-cost solutions without creating unnecessary administrative burden.

159 'Regulators Code' Department for Business, Innovation and Skills Better Regulation Delivery Office (2014). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/913510/14-705-regulators-code.pdf (accessed October 1, 2020).

“

Who gets to decide what is reasonable in a particular context? Who has the authority to do that?

Equality Task Force member

Annex 1: AI institutions

Introduction

This note provides a high-level overview of the key institutions in the UK that play a role in the way that artificial intelligence (AI) is used and regulated in the workplace. These institutions range from government bodies and regulators to non-profit organisations and academic organisations. There is no single regulatory body charged with overseeing the use of AI technologies in the workplace, so the mandates of the institutions listed below vary considerably – some are set up to facilitate a dialogue between the various stakeholders, whilst others determine if the use of AI breaches existing employment or discrimination laws or issue guidance.

This note also sets out some of their recent AI-related activities and the powers, if any, they have to investigate and/or penalise organisations that breach their rules. This note was kindly prepared by Freshfields Bruckhaus Deringer.

Government bodies

The AI Council

The AI Council is a non-statutory expert committee of independent members set up to provide advice to government and high-level leadership of the AI ecosystem. It was set up in May 2019 and aims to:

(a) enable the exchange of ideas between industry, academia and government, connecting and building on existing activities and ensuring fresh action where it is needed;

(b) advise across government on priorities, opportunities and challenges for the responsible adoption;

(c) share research and development expertise;

(d) horizon-scan for new AI technologies, applications and their impact; and

(e) work on improving the public perception of AI technologies.

The Government Office for AI

The Government Office for AI is part of the Department for Digital, Culture, Media & Sport and the Department for Business, Energy & Industrial Strategy. It works with government, industry and the non-profit sector and is responsible for overseeing the implementation of AI. Most recently, it issued Guidelines for AI procurement which summarises the best practices addressing specific challenges of acquiring AI in the public sector.[1]

Committee on Standards in Public Life (CSPL)

CSPL is an advisory non-departmental public body which advises the Prime Minister on ethical standards across the whole of public life in England. In February 2020, CSPL published a report on AI and its impact on public standards.[2] CSPL considers that data bias remains a serious concern and that there is an urgent need for practical guidance and

enforceable regulation. The report calls for the application of anti-discrimination laws to AI to be clarified. It suggests that the Equality and Human Rights Commission should develop guidance in partnership with the Alan Turing Institute and the CDEI (described below) on how public bodies should best comply with the Equality Act 2010 when developing AI.

Although the report identifies the need for a regulatory body to have responsibility for identifying gaps in the regulatory landscape, CSPL did not find a need for a new, separate AI regulator. Instead, it suggests that CDEI could be given an independent statutory footing to act as a central regulatory assurance body and advise the government and existing regulators on how to deal with AI issues. The idea behind this proposal is to allow existing regulators to continue to utilise their sector-specific experience while having the benefit of an expert regulatory body whose focus is exclusively on AI.

Centre for Data Ethics and Innovation (CDEI)

CDEI is part of the Department for Digital, Culture, Media & Sport. It is an independent advisory body set up and tasked by the government to investigate and advise on how best to maximise the benefits of AI technologies. It has a cross-sector remit and gives recommendations to regulators and the industry.

On 18 June 2020, CDEI published the AI Barometer,[3] an independent report which analyses the most pressing opportunities, risks and governance challenges associated with AI across five key sectors – criminal justice, financial services, health and social care, digital and social media and energy and utilities. Algorithmic bias leading to discrimination featured highly across almost all sectors. Over the coming months CDEI will promote the findings of the AI Barometer to policy-makers and other decision-makers from industry, regulation and research. The AI Barometer will be expanded to cover new sectors and gather more cross-sectoral insights.

Regulators

The Information Commissioner's Office (ICO)

The ICO is the independent body responsible for overseeing data protection in the UK, including the processing of personal data in the workplace. The ICO has been active in the field of AI through various initiatives, including:

(a) In collaboration with the Alan Turing Institute, the ICO has published guidance for organisations on how to best explain decisions made by AI systems to the individuals affected by them.[4] The guidance is not a statutory code: it was issued in response to the commitment in the Government's AI Sector Deal. Amongst other things, it includes practical tips for organisations on how to explain the steps taken to mitigate the risk of discrimination both in the production and implementation of an AI system and in the results it generates.

(b) Earlier in 2020, the ICO consulted on its draft guidance on the AI auditing framework.[5] The draft guidance includes recommendations for organisational and technical measures to mitigate the risks AI poses to individuals and provides a methodology to audit AI applications and ensure they process personal data fairly. Once in final form, the ICO will utilise this guidance in the exercise of its audit functions under the data protection legislation.

The ICO uses guidance and engagement to promote compliance by the organisations but if the rules are broken, organisations risk formal action, including mandatory audits, orders to cease the processing of personal data as well as monetary fines. For the most serious breaches, including failure to comply with any of the data protection principles, any data protection rights an individual may have or in relation to transfers of data to third countries, the ICO can impose fines of up to €20 million (or equivalent in sterling) or 4% of the total annual worldwide turnover in the preceding financial year, whichever is higher.

Annex 1: AI institutions

The Financial Conduct Authority (FCA)

In January 2020, the FCA and the Bank of England established the Financial Services AI Public Private Forum (AIPPF).[6] The purpose of the AIPPF is to promote constructive dialogue with the public and private sectors to better understand the use and impact of AI, including the potential benefits and constraints to deployment, as well as the associated risks. The AIPPF seeks to share information and understand the practical challenges of using AI within financial services, the barriers to deployment and potential risks, and to gather views on potential areas where principles, guidance or good practice examples could be helpful.

In a related development, in February 2020, the FCA announced a year-long collaboration with the Alan Turing Institute,[7] the focus of which is AI transparency in financial services. This project comes off the back of a survey[8] carried out by the FCA and the Bank of England and published in October 2019 which found that financial services are witnessing a rapidly growing interest in AI. The FCA has set out a proposed high-level framework for transparency in the context of AI which will be discussed at workshops with industry and civil stakeholders. The FCA and the Alan Turing Institute believe that there is no one-size-fits-all approach to transparency in the deployment of AI and that different stakeholders need to be considered independently and decisions on the information to be made accessible to them should be tailored based on a number of factors.

The Equality and Human Rights Commission (EHRC)

The EHRC is an independent, statutory public body and is responsible for enforcing the Equality Act 2010 and eliminating discrimination. Its approach to regulation varies considerably - from providing guidance and support for organisations in their compliance efforts to launching formal investigations, issuing compliance notices and bringing court actions. Breach of a notice or court order issued under the above powers can be enforced in court and lead to an unlimited fine. The EHRC remit and its regulatory powers enable it to intervene in situations where the use of AI in the workplace is challenged, especially in the context of algorithmic bias and discrimination. It could also collaborate with other institutions to produce guidance (see CSPL's suggestion above for a collaboration between the EHRC and the Alan Turing Institute).

The Health and Safety Executive (HSE)

The HSE is the UK's national regulator for workplace health and safety. It prevents work-related death, injury and ill health. It carries out inspections and investigates the most serious work-related incidents. The HSE also has wide enforcement powers, include the power to serve notices, withdraw approvals, issue cautions or prosecute duty holders. It also provides practical guidance and advice. The use of AI technologies can improve occupational health and safety but could also create risks, which is where the HSE could get involved by issuing guidance or inspecting how organisations comply with the health and safety rules when deploying AI.

Annex 1: AI institutions

Other institutions

The Employment Tribunal

The Employment Tribunal is an independent legal tribunal which makes decisions in legal disputes around employment law. It hears claims from individuals who consider that an employer or a potential employer has treated them unlawfully, including on the basis of discrimination. It can award compensation to successful claimants, which in discrimination cases is not subject to any monetary cap.

The Advisory, Conciliation and Arbitration Service (ACAS)

ACAS is an executive non-departmental public body, sponsored by the Department for Business, Energy & Industrial Strategy. It gives employees and employers free, impartial advice on workplace rights, rules and best practice. ACAS also offers dispute resolution services aiming to help employers and employees to reach an agreement without going to the Employment Tribunal. It publishes guidance on important issues for employers and although it has not done this to date, it could be involved in projects related to the use of AI in the workplace.

The Alan Turing Institute

The Alan Turing Institute is the national institute for data science and artificial intelligence. It is actively collaborating with various stakeholders both in the UK and abroad on various AI projects. As mentioned above, it has been supporting UK regulators – including the ICO and the FCA – in analysing the impacts of AI technologies and the development of guidance for organisations who wish to make use of these technologies.

Citizens Advice

Citizens Advice is a network of charities which offers confidential advice on various legal issues, including workplace discrimination. It has various tools and resources available to individuals to help them make an initial assessment of their situation. It has not been involved in the field of AI technologies and their use in the workplace yet, but its remit is wide enough to allow for this.

Conclusion

As mentioned above, there is no single institution with overall responsibility for regulating the use of AI technologies. The institutions listed above all have a role to play in this field but not all have done this yet. Only some of them have regulatory and enforcement powers allowing them to issue fines or take other measures against organisations who deploy AI in a way that breaches these institutions' rules or guidance.

Annex 1 endnotes

1 Guidelines for AI Procurement – 8 June 2020.

2 Artificial intelligence and public standards – February 2020.

3 CDEI: AI Barometer – 18 June 2020.

4 ICO: Explaining decisions made with AI – May 2020.

5 ICO: Guidance on the AI auditing framework – draft for consultation.

6 FCA announcement: Financial Services AI Public Private Forum.

7 FCA blogpost: AI transparency in financial services – why, what, who and when? – 19 February 2020.

8 FCA and Bank of England: Machine learning in UK financial services – October 2019.

“

From an equality law point of view, if an employer can’t explain why they recruited someone, that’s equivalent to saying ‘they’re just not my kind of person.

Equality Task Force member

Authors

Dr Reuben Binns (academic capacity)
Dr Abigail Gilbert
Dr Anne-Marie Imafidon MBE
Tim Johnston
Dr David Leslie
Joshua Simons
Helen Mountfield QC
Anna Thomas

Acknowledgements

Thank you to:
Paula Hagan
Samuel Atwell
Professor Jeremias Prasl
Andrew Burns QC
Ben Jaffey QC
Sa'ad Hossain QC
Dr Logan Graham
Andrew Pakes
Profesor Lilian Edwards
Carly Nyst
Bethan Chalmers
David Mendel

Equality Task Force

Members are:

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

Rebecca Thomas
Principal, Work 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 (ICO capacity)
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

Dr Logan Graham
Research Fellow (former), 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 Equality Task Force for their time and contributions.

Institute *for the* Future of Work

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

www.ifow.org
@FutureWorkInst

Registered charity number 1180718