ENAR'S ELEVENTH EUROPEAN EQUAL@WORK SEMINAR

TOOLKIT

ARTIFICIAL INTELLIGENCE IN HR

HOW TO ADDRESS RACIAL BIASES AND ALGORITHMIC DISCRIMINATION IN HR?

european network against racism

Published by the European Network Against Racism (ENAR) in 2020 in Brussels, with the support of the Rights, Equality and Citizenship Programme of the European Union, the Open Society Foundations, the Joseph Rowntree Charitable Trust, the Sigrid Rausing Trust, the ENAR Foundation and ENAR's business partners: Adecco Group, Coca-Cola European Partners, Inditex, L'Oréal, Sodexo and Twitter, as well as Google.

The contents of this publication can in no way be taken to reflect the views or opinions of the European Union, the Open Society Foundations, the Joseph Rowntree Charitable Trust, the Sigrid Rausing Trust, the ENAR Foundation, Adecco Group, Coca-Cola European Partners, Inditex, L'Oréal, Sodexo, Google, or Twitter.

Acknowledgements
This toolkit was drafted by Dr Lucy Michael and Dr Bethany Waterhouse-Bradley.

ENAR thanks the speakers and participants who took part in the 11th Equal@work seminar on AI, algorithms and recruitment, on 6 December 2019, for their valuable contributions which have been used for this toolkit.
Design and layout: Crossmark

ENAR – European Network Against Racism aisbl
Tel +32 2 229 35 70
Email: info@enar-eu.org
www.enar-eu.org

Table of contents

Introduction

Since 2015, businesses have increasingly adopted Artificial Intelligence (AI) solutions with the view to improving a wide range of standard human resources functions; from candidate engagement, hiring and promotion to disciplinary procedures and terminations. In many cases, these tools act as more streamlined, higher capacity versions of existing practices (e.g. sorting and classification, determination of eligibility, and assessing 'cultural fit'). AI can also be used to assess employee satisfaction and improve retention by providing a tool which can manage and analyse data at a scale not previously possible. Employers seeking to ensure access to the most suitable candidates and to transform HR processes from largely operational to strategic are most likely to be motivated learners in this area, but the broad marketing of AI-based solutions for HR makes this knowledge important to a much wider audience of HR and Diversity & Inclusion Managers.

As jobseekers increasingly have access to online job search services which allows them to produce multiple applications with limited effort, HR teams are often left sorting through unprecedented numbers of applications for a single post. The streamlining of various aspects of hiring processes, as well as managing increasingly remote and disparate teams, is enticing in what is becoming a highly digitalised labour market. The use of data-driven technology in recruitment is predicted to be a growing trend over the next few years. There is increasing reliance on automation either to supplement or replace a series of tasks in the human resources sphere. These technologies could range from relatively simple text scanning technology, to more complex content analysis and even artificial intelligence-led interviews.

The use of algorithmic decision making can reduce time for HR professionals and hiring managers in screening large numbers of applicants, improve selection processes, and provide the potential for predictive analysis. Recruitment is the principal arena in which AI functions are being adopted. This includes targeted recruitment advertising, bulk screening of CVs and applications, providing recommendations to human decision makers on who to invite to interview and analysing candidates' performance in selection tests and interviews. The high costs of recruitment and training make rapid and large-scale data-informed decision making a crucial part of the competitive business landscape. This in turn helps businesses to reduce the costs associated with staff attrition and decreased performance, and increase the strength of long-term workforce planning, making the HR function more strategic and capable of improving decision making in other parts of the business too.

As awareness grows of the problem of racist attitudes in recruitment – particularly with the popularity of the 'unconscious bias' approach and growing market of solutions to address it – many employers assume that automated decision making is more effective at reducing bias than human hiring managers. Some vendors of AI also offer products which are specifically designed to mitigate against human bias in selection, reinforcing the view that AI is at worst neutral on the issue. However, there have been notable problems with AI in this area which illustrate that bias is built into algorithms, consciously and unconsciously. Amazon, for example, withdrew its recruitment algorithm when it was discovered to both reproduce and amplify existing gender bias in the technology sector. **Without a clear view of the problem, algorithmic technology is currently more likely to further embed, rather than disrupt, biases which have excluded whole categories of applicants from employment in certain sectors and roles.**

AI is now recognised to reproduce and amplify human biases, and the particular capacity for this to exaggerate bias in HR processes is widely acknowledged as deserving of attention. Algorithms can reinforce discrimination if they focus on qualities or markers associated only with particular (already dominant) groups. While some of these markers are easily recognised (e.g. career gaps and gender), the current lack of racial diversity in workplaces across Europe makes markers of racial bias less well recognised. Algorithms which reinforce these biases will further reduce diversity across the European labour market and reduce corporate flexibility in long-term workforce planning as well as the development of markets. In spite of awareness of these emerging problems, the scale of the potential issues is often understated, and the focus often on the technical aspects; or fixing the tools. This serves to both increase the likelihood of these effects being overlooked, and to shift focus away from the structural and institutional problems which often lead to the production of undesirable outcomes from the use of AI.

The specific causes of algorithmic bias can be difficult to detect after the fact, making local expertise and transparency key for trust and accountability (by management as well as employees), and centring the need for upskilling, expert collaboration, safeguards and explicit policies for algorithmic development and design. **Crucial to the mitigation of risks associated with unethical or biased intelligent systems is an increased understanding of how algorithms are designed, trained, and the context in which they are utilised.**

On a societal level, the systemic exclusion or over-representation of certain groups in certain professions can embed social inequalities. The potential for intelligent algorithmic systems to improve current job matching services is significant. Being able to recommend jobs to people that they might not search for or think themselves able to apply for is a development that many in the recruitment world are now exploring. Encouraging the development of these systems could benefit many people but recruiters will need to ensure that their recommendations are not discriminatory. There are calls for increasing public oversight of the use of AI tools[1] but the extension of these to the research and development processes of these tools to ensure compliance with human rights standards increases the likelihood of meeting the standards of transparency, human oversight and robustness set out in the 2019 EU White Paper on Ethical AI from the early stages of development. This is particularly relevant where companies are developing tools in house. **Public oversight and application of regulatory frameworks, paired with widening participation in the development of AI tools, have the scope to mitigate or reduce bias in their use.**

Diversity in the workforce is key to business success today. There is a direct correlation between the diversity of a workforce and the breadth of its perspective.[2] Diverse workforces are also more productive, so employers should actively seek ways to recruit candidates into the workforce from different backgrounds. But trust in algorithmic decision making can be decreased by poor recruitment outcomes, public accountability for bias, and errors in selection and performance functions, leading to 'algorithmic aversion'. Building confidence amongst HR and D&I Managers that AI can reduce risks for firms in respect of discrimination and recruitment costs is a valuable service to business.

Outline of toolkit

This toolkit is designed for Human Resources and Diversity & Inclusion Managers, as well as Programmers, to ensure that consumers of off the shelf and custom AI solutions for Human Resource Management have a clear guide to challenges, solutions and good practice, in a format which supports conversations with Programmers providing solutions.

The toolkit explores the role of human bias and structural discrimination in discriminatory or unethical AI programmes, and provides clear and practical steps to ensure companies have the necessary cultural and technological tools to responsibly digitalise HR systems with the help of intelligent systems. In doing so, HR and D&I managers will come to understand bias reproduction and amplification, and gain the confidence to address bias risks produced by inadequate or inappropriate training data, simplistic or reductive classifications or other human errors leading to biased outcomes of algorithmic decision making. Importantly, it will support HR teams in effectively transferring existing knowledge of discriminatory hiring practices and building diverse workplaces to the responsible deployment of intelligent systems to aid in those objectives.

About the Equal@work Platform

ENAR considers that the best way to achieve substantive equality is through collaboration and dialogue between different actors – private companies, public administrations, trade unions, NGOs, employees – to find solutions or share good practices.

The Equal@work Platform is a space for employers, trade unions, public authorities and NGOS to collaborate for innovative solutions to diversity management. Members of the platform explore how to integrate an anti-racist approach; ensuring improved access to the workplace for people of colour and an end to structural discrimination in the labour market.

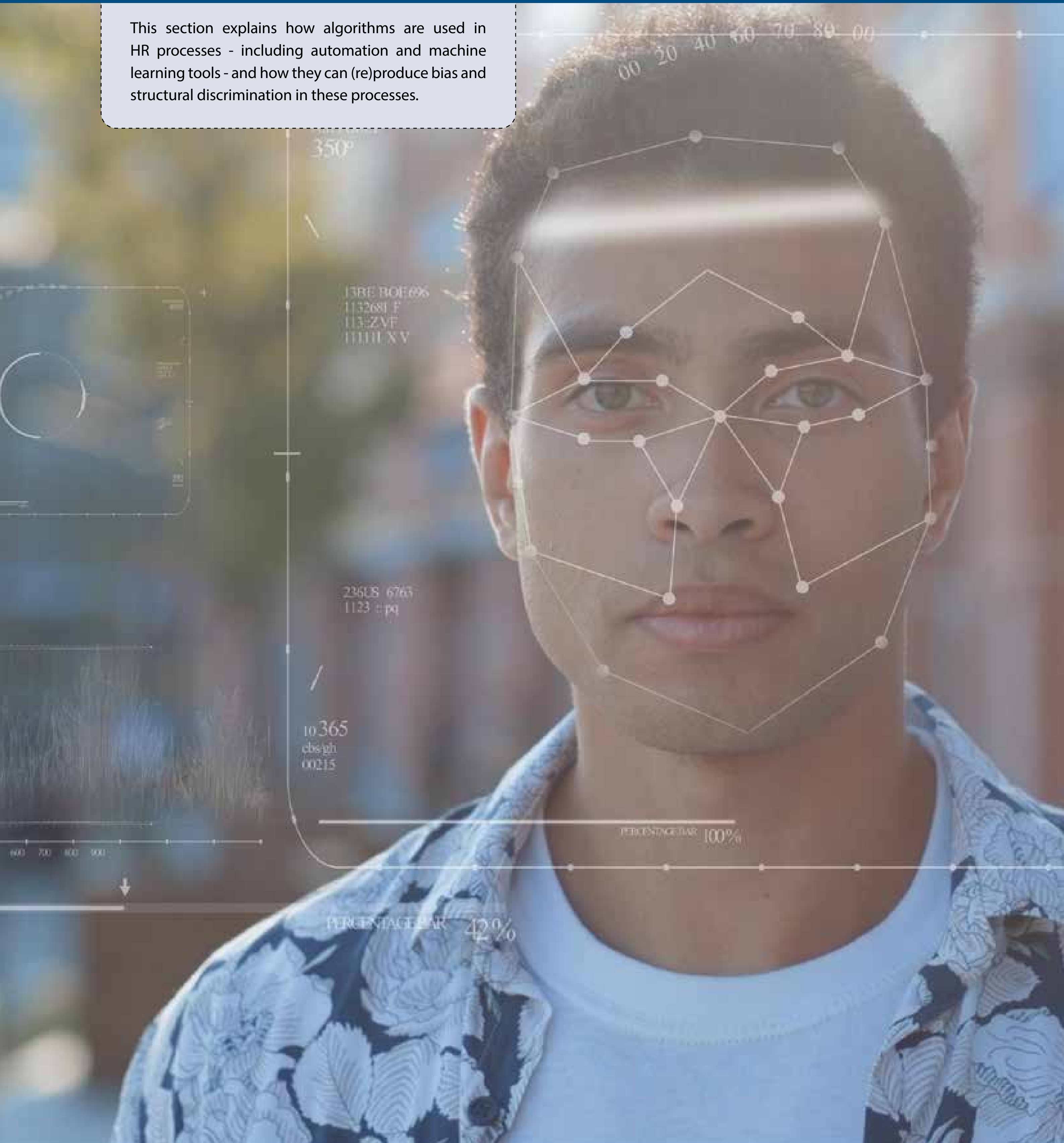
This toolkit was produced as a follow-up to the 11th Equal@work seminar on AI, algorithms and recruitment, organised by the European Network Against Racism in December 2019 with the support of Google.

[1] Engler, A. 2020. 'The European Commission considers new regulations and enforcement for "high-risk" AI'. [blog] Brookings Institute, 26 February. Available at: https://www.brookings.edu/blog/techtank/2020/02/26/the-european-commission-considers-new-regulations-and-enforcement-for-high-risk-ai/.

[2] Slater, S. F., Weigand, R. A., & Zwirlein, T. J. 2008. 'The Business Case for Commitment to Diversity'. *Business Horizons, 51*(3), 201-209; Catalyst. 2020. *Why Diversity and Inclusion Matter: Financial Performance.*

1. UNDERSTANDING ALGORITHMS IN HR

This section explains how algorithms are used in HR processes - including automation and machine learning tools - and how they can (re)produce bias and structural discrimination in these processes.



Although there is no single agreed definition of AI, there are similarities between many of those being used. Broadly, AI is “a set of statistical tools and algorithms that combine to form, in part, intelligent software” enabling “computers to simulate elements of human behaviour such as learning, reasoning and classification”. Often confused with AI, ‘machine learning’ algorithms are a narrower subset of this technology. They describe “a family of techniques that allow computers to learn directly from examples, data and experience, finding rules or patterns that a human programmer did not explicitly specify”. In contrast to conventional algorithms which are fully coded, the only instructions given to machine learning algorithms are in its objectives. How it completes these are left to its own learning.

Machine learning is employed to categorise (people, behaviours or things), predict (behaviours, outcomes, actions), identify (patterns, relationships), and detect (risks, anomalies) using algorithms. While machine learning does have the ability to independently perform tasks, it does so using directives from human designers and is trained using data points which are collected through human-led processes and procedures. This means that it is subject to the same biases and prospects of discrimination that human HR systems would be. Machine learning is most usefully employed for tasks where there is a straightforward, easily classified and categorised answer which does not require nuance or interpretation, or where the data is so plentiful and complex, a normal system or person could not process it.

The multiple stages of the hiring process, which typically consist of various forms of screening and elimination, include many tasks which are automatable such as: online targeting, CV screening, correspondence with candidates, tracking communication between company and client, career progression and workforce development. Email screening, video interview analysis, predictive analytics are used to infer individuals whom organisations might want to recruit, and target online job advertisements at them, based on users’ search histories, usage patterns and demographics. Employers may utilise predictive resume review software to automatically scan resumes for certain keywords and rank applicants based on their predicted suitability for the position. After a candidate’s resume is selected for further review, companies can use third-party interview analysis software to automatically score interviewees’ facial expressions, choice of vocabulary and tone. Companies can even use algorithmic assessments to predict a candidate’s job performance before they even step into the office, including their likely progression through several roles.

Unlike common perceptions of total autonomy and objectivity, machine learning tools operate without human oversight, but work based on pre-programmed commands. The machine then uses data to refine and adapt the way they complete the tasks. It can range from working based on commands derived entirely by humans, or learning and adapting from those commands (supervised or semi-supervised), to replicating or reinforcing human work, the machine, or algorithm, to autonomous working through using data sets to independently identify patterns and relationships, adapting the algorithm as it continues to learn. These are often referred to as automated tools. Deep learning takes autonomous working to the next level. It analyses more complex data at a higher volume. Instead of simply identifying patterns, it attempts to intuit from and mimic human behaviours. These are often referred to as predictive tools. **Given the interdependence of AI on human input and data filtered or generated by humans, it is clearly vulnerable to the same individual and structural biases which can prove challenging in fair hiring and employment.**

Automation tools create faster ways of identifying and categorising data based on a set of established criteria. Predictive tools aim to forecast outcomes and behaviour by analysing existing data. Predictive features rely on machine learning techniques, where existing data is used to train computers to detect patterns in that data, and build models that forecast future outcomes. Scores and rankings are assigned to candidates that are deemed most likely to be successful. While these tools are often employed to supplement human decision making, where the employee and the machine disagree, there is a tendency to trust the computer over the human.

It is not uncommon for both employers and applicants to be unfamiliar with how these algorithms – and especially unsupervised learning models – work. With the exception of large technology companies like Google and Amazon, who have the capacity to develop intelligent systems in-house, most industries rely on external providers to either facilitate an automated service, or to design a bespoke tool. Third-party algorithms are almost universally opaque due to factors like proprietary software, patented technologies, and/or general complexity. Non-developers can be unaware of which specific factors motivate algorithmic rankings, as well as what training data and de-biasing techniques are utilised. Additionally, non-developers (in this case, most hiring managers) often lack the ability to modify the third-party algorithms that they use. Because of this possible disconnect, employer intent is a relevant consideration when judging potential discrimination in algorithmic software. Those who are able to afford bespoke tools must also look at the sustainability of these machines, which can require extremely frequent updating as well as access to powerful computers capable of hosting the algorithms.

As such, most employers are more likely to employ an 'off the rack' service, such as Workable or HireVue. Workable claims to have supported more than 20,000 companies as of September 2020, and HireVue reported collecting data from more than 12,000 clients as early as 2015.

1.1. What do we mean by bias in this context?

Hiring biases are produced partly by unconscious bias, which favour some groups more than others. But conscious biases, for particular types of career history, education, areas of interest, play a large part in who we select. These operate on top of structural racism which affects minority candidates' access to educational institutions, work experience, employment histories, pay and promotion, and institutional racism which affects how your organisation interacts with minority ethnic clients or candidates, and narrows the potential candidate pool. Protecting your organisation against racial bias is good for diversity in the workforce, but also is a proven strategy to promote innovation, creativity and productivity in your business. The economic and workplace advantages to having more heterogeneous personnel include improved client relationships and better decision making. While there is some indication of positive changes being made in terms of overt gender and race discrimination, breaking down the analysis will show that women and people of colour are still severely underrepresented in CEO positions, STEM fields and senior medical professions. Women continue to be paid significantly less than their male counterparts; a statistic that becomes exponentially worse if you are a woman of colour. Changes in diversity do occur but tend to be isolated to specific areas or minority groups or limited to the lower organisational levels of high status professions. It is worth noting that the profile of the tech sector, including external providers of algorithm-based HR tools, is predominantly white, male and young, and that stemming widening racial and gender disparities has been challenging due to discriminatory workplace cultures.[3]

Interpersonal racism is the bias that occurs when individuals interact with others and their personal racial beliefs affect their public interactions. This can be expressed as discrimination, mistreatment (including microaggressions) or violence. They can also be apparent in treatment of clients, service users, suppliers or other people in contact with an organisation. These forms of racism are most often addressed through grievance procedures.

Institutional racism describes the way in which the policies and practices of an organisation, which advantage particular groups over others. It can be seen or detected in processes, attitudes and behaviours that amount to discrimination, either through unwitting prejudice, ignorance, or thoughtlessness and racist stereotyping, which disadvantage ethnic minority people.[4] This results in the services, resources and opportunities of an organisation being unavailable to others, routinely producing racially inequitable outcomes. In recruitment, this can mean the design of job opportunities for certain groups, based on irrelevant characteristics. Searches for candidates who will 'fit' the organisation often prioritise culture over skills and experience, and reproduce a homogenous workforce with low capacity to ensure the organisation thrives in a diverse society. Anti-discrimination laws and training operate best at this level, although the unconscious nature of much institutional racism requires active and explicit measures to address it. Institutional racism can even affect the way in which complaints about racism are dealt with, exposing complainants to harassment, exclusion and loss of resources and opportunities including employment.

Structural racism is a product of a system in which public policies, institutional practices, cultural representations, and other norms work in various, often reinforcing ways to perpetuate racial group inequity and has been a feature of the social, economic and political systems in which for instance we all exist.[5] This refers to the ways in which education, residence, work experience, wealth and social networks have reinforcing effects to provide much greater life chances to some groups. Importantly, structural racism appears neutral, but constantly adapts to new circumstances to preserve advantage for dominant groups. Recruitment which values particular universities, for example, reproduce the discrimination in education and university application systems reflecting wealth, language, social networks, behaviours, subject choices, and extracurricular activities.

Systemic racism underpins both structural and institutional racism. The historical position of a highly advantaged group determines long-term power over institutions, policies and law and establishes what is considered 'normal' or desirable, by playing a key role in setting standards for culturally valued behaviours, language and knowledge. Under systemic racism, systems of education, government and the media celebrate and reward some cultures over others. When systemic injustices remain unspoken or accepted,

[3] Kapor Center for Social Impact. 2017. Tech Leavers Study. Available at: https://www.kaporcenter.org/tech-leavers/.

[4] Report of the Stephen Lawrence Inquiry. 1999. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/277111/4262.pdf.

[5] with this text: Aspen Institute. Structural Racism and Community Building. 2004. Available at: http://www.aspeninstitute.org/sites/default/files/content/docs/rcc/aspen_structural_racism2.pdf.

an unethical white privilege is fostered. Few laws or policies are aimed at explicitly addressing systemic racism but understanding it is key to actively addressing its effects in your organisation.

When a person is disadvantaged by virtue of multiple identity categories (race, gender, age, disability, sexuality, nationality, etc.), that disadvantage is not simply additional, but multiplied. This is described as the **intersectional** experience. This approach takes into account people's overlapping identities and experiences in order to understand the complexity of the discrimination and inequalities they face.

1.2. How do algorithms (re)produce bias?

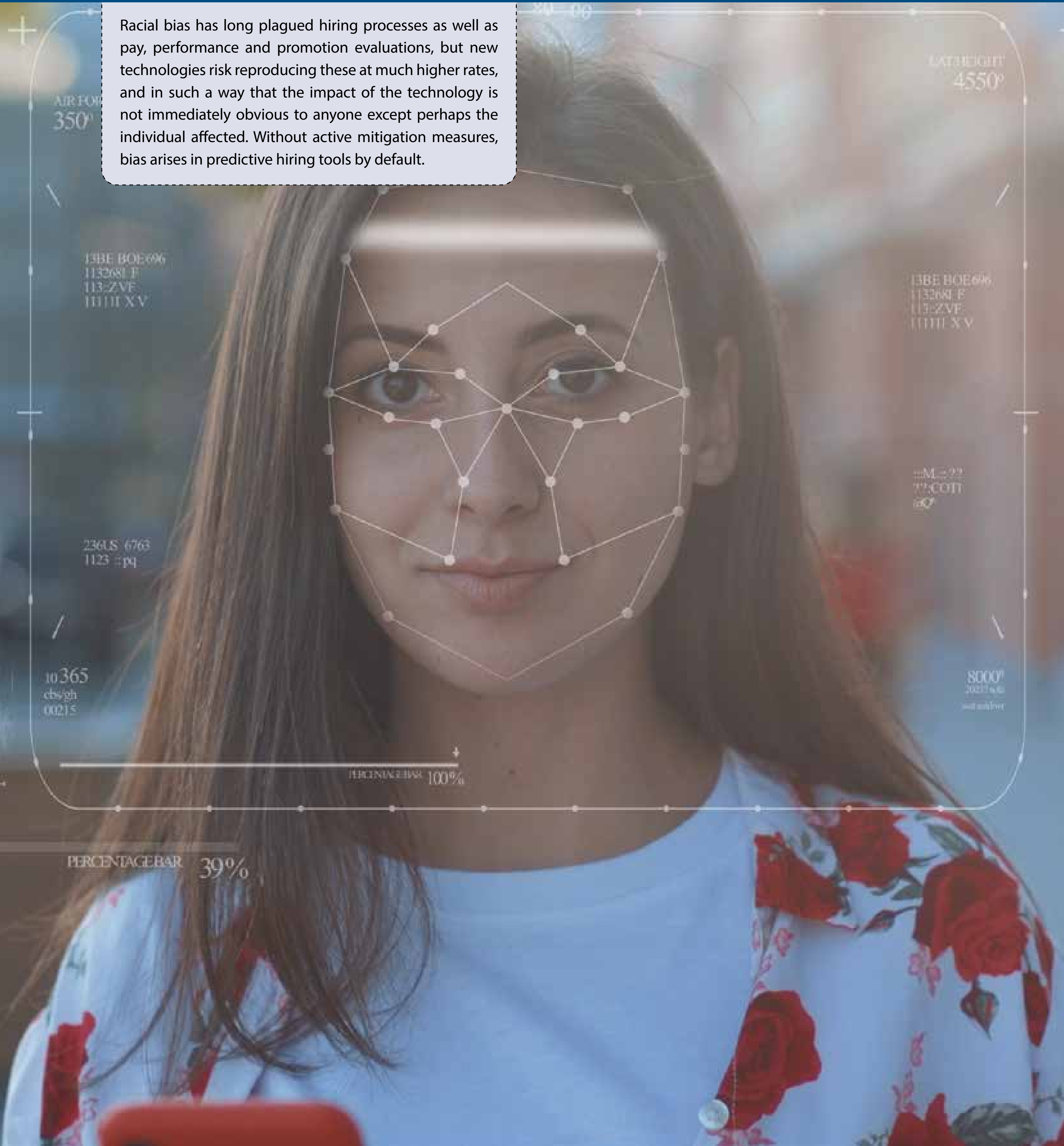
Algorithms are used to help recruiters and employers make more informed decisions. When tested, retested and implemented effectively, there is scope for algorithms to mitigate the limitations of human unconscious bias and open up a talent pool to increase diversity in the workforce and attract greater talent. Tools can be retrained and improved – faster and more effectively than changing human perceptions. However, **the assumption that algorithms are neutral by nature is mistaken**. Algorithms are trained directly on the brief provided by clients. While they analyse complex information, the objectives of algorithms must be relatively simple. Without care and attention in commission and design, these sometimes very literal interpretations of employer briefs are more likely to reproduce and scale up prejudices than to defend against racial bias in the first instance. The perception of neutrality, however, leads to employers being overconfident in adopting intelligent systems and underprepared to tackle the discrimination which results. Bias hidden in the algorithm is much more difficult to recognise unless we build our capability to do so.

Additionally, there is no standard definition of bias amongst the designers of intelligent systems. Designers rely heavily on cognitive processes in the design of AI, and so they tend to understand bias in psychological terms. **A focus on unconscious bias means that there is little attention paid to the much larger problem of structural discrimination**, which produces accumulated disadvantage through blocked access to key institutions and opportunities. This is embedded in educational, employment and residential segregation, and experiences of illness, care and economic insecurity, as well as fewer connections with groups who have unfettered access to institutions and fewer experiences of insecurity and exclusion produced by structural discrimination.

Algorithmic bias can also exaggerate the power imbalance between employers and candidates or employees. Candidates rarely understand how algorithms work to shape and filter their application experience and ultimate success, particularly because few employers are transparent about how they are constructed and used. Machine learning relies on huge numbers of data points. HireVue is purported to have the ability to collect more than 50,000 data points in one video interview. It is not clear how much potential candidates have understood or consented to the use of their data in specific ways. The leads to risk of violating GDPR, obfuscating discrimination and reinforcing information asymmetry between employers and candidates/employees.

2. MACRO CHALLENGES:
HOW STRUCTURAL RACISM IS REPRODUCED AND AMPLIFIED BY INTELLIGENT SYSTEMS

Racial bias has long plagued hiring processes as well as pay, performance and promotion evaluations, but new technologies risk reproducing these at much higher rates, and in such a way that the impact of the technology is not immediately obvious to anyone except perhaps the individual affected. Without active mitigation measures, bias arises in predictive hiring tools by default.



2.1. Reproducing existing bias

Machine learning tools tasked with automation and prediction rely on existing data sets to 'train' the algorithm. This means that data from the existing workforce (be that industry or specific company) is the basis for the measures used to determine whether a candidate is worth hiring or not. In this way, the individual bias of those involved in hiring practices and development of the tool are transferred to the algorithm. This is a case of a biased algorithm, but the problem itself is not a technical one, but rather the lack of diversity in the data pool which is often a direct result of biased hiring and employment practices. This is of particular concern for international application of software, which is likely to reproduce characteristics which have the scope to homogenise lists of desired language, qualifications, etc. In this way, the existing culture is reproduced. Vendors' claims that their tools will naturally reduce bias by obscuring applicant's sensitive characteristics refer only to the risk of interpersonal discrimination by a recruiter. Other forms of discrimination, which are institutional or structural, are not addressed by the same means. These are critical when it comes to reducing the amount of bias in hiring or HR processes, but are precisely the types of bias which technology reproduces and amplifies at speed. **Trained on the existing workforce and performance benchmarks determined from the successes and perceived failures of those who already work for the organisation, new hires continue to resemble those hired before** (as they are all based on the same characteristics and means of portraying those characteristics).

This is an example of how the technology begins to shape the human after the human has shaped the technology. People are being trained on how to interview in a way that is well received by the system. Because machine learning tools continue to adapt and change to become more efficient, even if they are given criteria which excludes sensitive characteristics (e.g. race, gender), it is possible for that tool to begin to capture measures which are soft indicators of cultural norms. This was seen in the inclusion of the name 'Jared' and a hobby of lacrosse as the two most important criteria in inviting applicants for interview when auditing a CV screening algorithm. These soft indicators are strongly correlated with white, middle class males and are representative of the homogeneity of the employee pool on which they were trained. Amazon faced similar scrutiny when women's sport was used as a criterion to eliminate applicants. 'Cultural fit' indicators, such as hobbies or interests, are often strongly correlated with certain socio-economic classes, nationalities or racial groups. By reliance only on data about people who are perceived as 'high performers', or 'long stayers' within organisations (the latter being impacted by issues of racial harassment, exclusion or return-to-work discrimination against women, for example),[6] automation and predictive tools also run the risk of reproducing these biases.

Targeted recruitment tools often reproduce inequality through the types of positions and level of opportunities they present to certain groups. Recruiting candidates through online advertising will shape who sees job adverts. Historically this has led to incidents where, for example, female applicants are less likely to see high paying opportunities. It also relates directly to who has active and dynamic online profiles. Ethnic minority candidates, especially female, are less likely to have open social media profiles due to reporting higher rates of online abuse. The differentiation in methods of recruitment of high level executives – through agencies rather than online tools – means that the issues with bias reproduction through automated recruitment are once again targeted at the groups most likely to experience discrimination in hiring.

2.2. Lack of transparency and accountability

An absence of firmly established governance and legal frameworks means it has been very difficult to enforce the accountability and transparency of predictive analytics tools. However, there are rapidly emerging discussions of how existing frameworks on non-discrimination, data protection, digital rights and employment regulations and their interoperability should adapt to algorithmic decision making, and rapid policy development is under way. **The opacity of algorithmic decision making, largely because of proprietary rights of the vendors, has made it difficult to increase transparency.** However, this is likely to be driven by consumer demands, and public sector contracts increasingly seek transparency around the algorithms commissioned.

Individual discrimination cases have been undermined by the lack of feedback from automation and predictive systems, with lack of evidence because of proprietary protection of algorithms, but internationally courts are seeking ways to address the increasingly evident harms which result from bias in these systems. New techniques are being developed to establish accountability, such as the requirement to produce counterfactual accounts. These give candidates information about the benchmarks applied, and the gap between a successful and unsuccessful application.

[6] O'Brien, K.R., Scheffer, M., van Nes, E.H., van der Lee, R. 2015. How to Break the Cycle of Low Workforce Diversity: A Model for Change. *PLoS ONE 10(7)*. Available at: https://doi.org/10.1371/journal.pone.0133208.

Public sector organisations have been at the forefront of discussions about trust in algorithmic decision making. Organisations like Nesta in the United Kingdom have sought to support the use of AI with guides to aid transparent introduction, while the New Zealand Government has introduced a Charter for all government agencies, which require them to assess the likelihood of bias and document the risks arising in use.[7]

2.3. Reliance on simplistic classification

When defining and comparing groups, we rely on clear definitions of groups based on ethnicity, age etc. and external definitions of one's ethnicity. When a system differentiates between groups of people for a particular purpose based on their gender, ethnicity, age, dis/ability, or socioeconomic status, it gives them particularly meaning. While the experiences of people in these categories are shaped by various forms of discrimination in each society, those categories themselves are constructed in particular contexts. For this reason, descriptions of racial or ethnic groups adopted in the United States, for example, have different outcomes when applied in Europe.

Algorithmic systems frequently model those categories as fundamental attributes of people. In an attempt to increase the 'elegance' of an algorithm, categories such as gender, race, ethnicity and disability must be simplified. **This process could lead to the exclusion or alienation of those whose identities are poorly served by a lack of nuance and complexity in definition** (e.g. individuals who are transgender or non-binary, mixed race or with dual nationality), and force conformity to definitions which fit within predominant categories. Systems which rely on yes/no categorisations also have more difficulties comparing intersectional experiences (e.g. white man versus black disabled woman).

2.4. Problematic measures

Increasingly, the employment of psychometric tests, IQ measures and generally incorporating personality types and traits into recruitment has been shown to be problematic at best and discriminatory at worst. There are examples of AI being employed to administer these tests, and/or being developed using the type of measures used in these tests. This is another example of AI being used to perpetuate or reproduce existing discriminatory measures, however because it is part of a technical system, it is more likely to be deemed neutral or objective. AI can use analysis of facial movements and speech patterns to judge the level of enthusiasm and preparedness for the job. While it is true that in many cases AI systems have been deemed to be more accurate than humans in object recognition, this is only the case for the 'best AI systems', which is likely to require equipment and machines which are only accessible to some of the largest employers. There are technological challenges in facial recognition, language recognition and usage, as well as problems of interpretation across culture, class and race. It is necessary to understand the limits of this type of technology at present, and the extent of work necessary to make these transparent.

2.5. Imbalance of information and capacity

AI and Machine Learning are complicated ideas, and most people outside of the technology sector find the idea of trying to better understand how they work daunting. The 'charisma of numbers' can often give false confidence about certain outcomes. This has been evident for many years with the employment of traditional statistical techniques, but becomes even more inaccessible when AI is introduced. This poses a problem of capacity within organisations, where it is likely that the people using the tools have a poor level of understanding of how they work, and importantly, how they might go wrong. Development of software is based on a brief provided by your organisation. Consumers therefore need to understand how to write a brief for AI designers, especially since interpretation tends to be literal.

There is also an imbalance of information between an organisation and their employees or candidates in terms of data collection and usage in the development and application of algorithms. In video interviews, for example, Hirevue claims that a 30-minute session can 'yield up to 50,000 data points'. The reduction of candidates to these data points, the issues associated with how the data points are chosen, how they are weighted and valued, are all things that should be screened by the employer, as well as ensuring that the user has an awareness of the data points being collected.

There are a range of risks to existing members of the workforce; both in the company more broadly, and in the HR departments. First, there is the risk of replacement and the resultant unemployment/underemployment which could arise as administrative HR tasks are increasingly being carried out by systems. This can be worsened by a capacity loss

[7] The Algorithm Charter for Aotearoa New Zealand. Available at: https://data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-2020_Final-English-1.pdf.

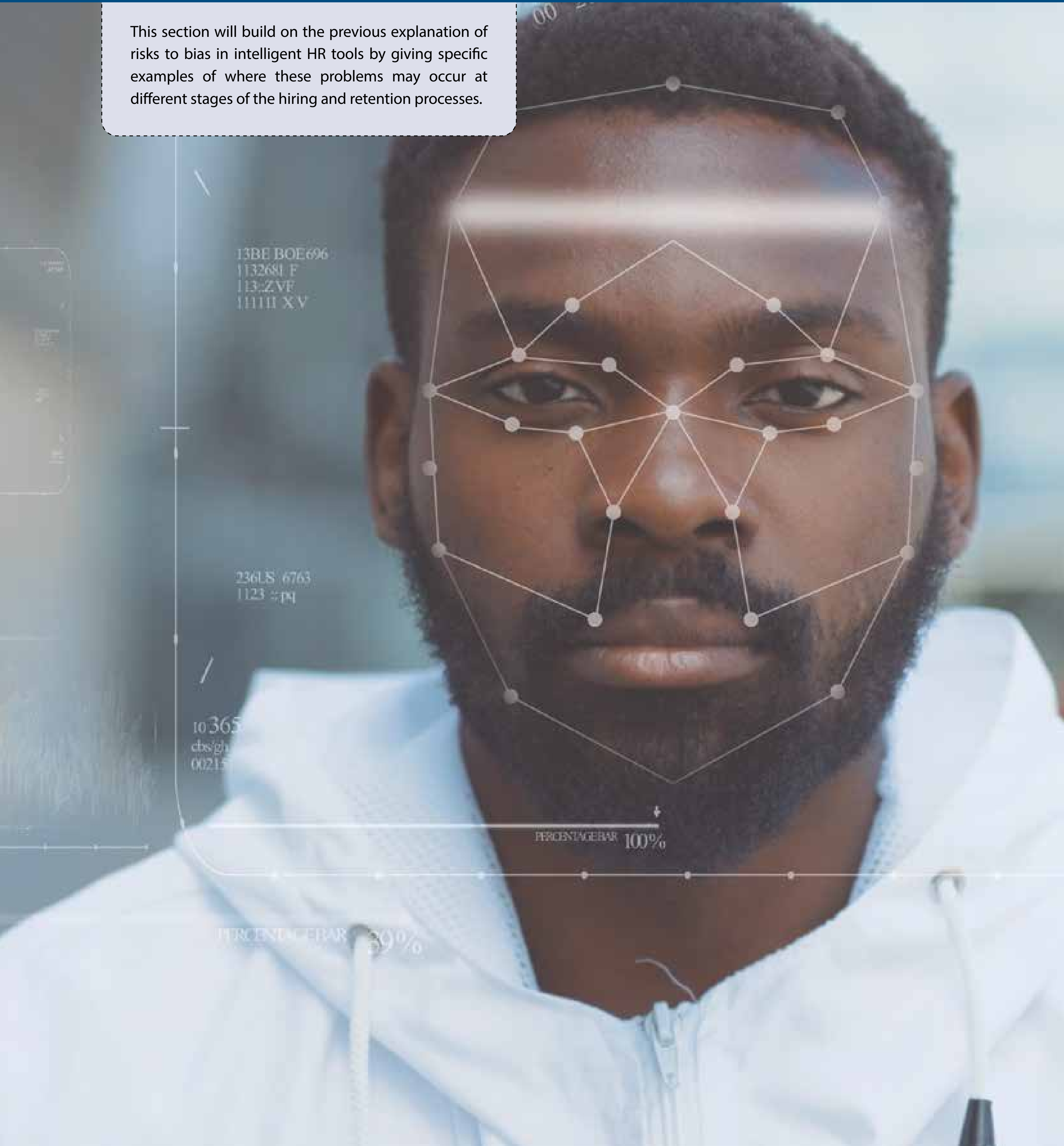
of those staff members who do remain, as they are limited in the scope and frequency of some of the most basic tasks in the profession. Generating new dependence on machines without sustainability models could lead to problems arising from technical failures or system shutdowns.

2.6. Informed consent and opt-out

There is a need to carefully consider the issue of informed consent. This has two key components: firstly, do people have enough information about how their data will be used and stored (including issues of confidentiality and anonymity); and second, can we consider this genuine consent when it is a requirement to be considered for the post? This is particularly important as the use of AI in employment becomes more prolific. While many HR teams want to employ AI tools to measure the successes and strengths of employees, there are a number of examples in the private sector of this same data collection being used to sanction staff and evidence terminations. Once again, the issue of consent and data protections comes into play, this time for existing employees. Can staff really 'opt out' of a system on which career progression depends? Data volume and variety is key to accuracy and generalisability of results. Complex data, or data with 'high dimensionality', is better for training AI, but is often more personal in nature and relates to audio files, video and photographic images.

3. HOW BIAS IS REPRODUCED IN EACH OF THE KEY TECHNOLOGIES

This section will build on the previous explanation of risks to bias in intelligent HR tools by giving specific examples of where these problems may occur at different stages of the hiring and retention processes.



Algorithms are trained to aid employers in effective (and often more rapid) decision making at a variety of stages in the recruitment process and in human resources management:

- **Candidate engagement / targeted recruitment**
- **Screening (customisable and pre-built assessment)**
- **Performance evaluations (probation and interval)**
- **Employee satisfaction and attrition**

3.1. Targeted job advertisements

The first stage at which digital technologies are used is in generating (or sourcing) a strong set of applicants. This can be done via advertisements, active headhunting or attractive job descriptions. Digital advertising and personalised job boards proactively shape applicant pools. Job adverts can be optimised and their wording augmented to be attractive to selected types of applicants. The characteristics of the applicants may be based on those who have been successful applicants previously in the sector. While targeting can improve the reach and influence of job adverts to potential employees, it also reduces the diversity of the potential pool, ensuring that some demographics are less aware of opportunities, not just at a single firm, but across whole sectors. Candidates are often not aware of the extent to which this shapes their application history.

Discriminatory practices through online targeting can be both intentional and unintentional. **Where attributes act as a proxy for traits such as race or gender, your targeted job adverts are already eliminating potentially strong applicants before application**. But even with open or 'inclusive' targeting parameters being set, advert delivery can still end up being unintentionally skewed across racial and gender lines by, for example, making it more expensive to target adverts at some groups of job candidates. This is described as *discrimination through optimisation*.

3.2. Screening applicants

The main function of digital and predictive hiring tools is elimination of candidates, rather than selection for appointment. Since much of this activity occurs early in the hiring process, the highest rate of bias occurs at this early stage. This is where applications are eliminated for not meeting the minimum or desired criteria to progress in the application process. **This is where algorithmic bias can strongly influence whether an application is rejected**.

Screening uses algorithms that systematically decipher a cover letter and CV and save this information in the company's HR database. This information could include years of experience, the languages spoken, qualifications obtained and the countries in which a candidate has worked. Algorithms are used to narrow down the selection of candidates automatically – not in an affirmative way, but by rejecting those who do not fit. Machine learning algorithms are used to screen resumes for keywords in context and to create relative rankings between the different candidates.

Collection of biometrics – physiological characteristics relating to body, facial recognition, DNA, hand geometry, iris recognition, micro expressions, odour/scent and retina scanning – is used mainly for authentication of the candidate's identity. It is also utilised for typing rhythm, gait and voice patterns. It is difficult to understand the reasoning behind many of these inclusions, through which the invasion of personal privacy and risk of reproducing structural biases outweighs the rather unclear benefits of these authentication measures.

Automation of the screening process relies on traditional (and often structurally biased) factors like qualifications, basic demographics, work experience, longevity in employment and geographic location. Predictive technologies however can work in several ways; using chatbots and virtual interviews to engage candidates, or even using game-based assessments. These different approaches both reproduce racial bias, albeit in different ways, and the remedies available differ because of the ways that the technology assigns value in its appraisals. We will outline some of the predominant examples of this in the next section.

3.3. Chatbots and virtual interviews

Chatbots are used to give information to applicants, to screen candidates and arrange interviews or further assessments. Chatbots use sophisticated algorithms that learn by millions of examples. Such magnitude of language data only exists in public data sources like news articles and social media. However, to deal with inadvertent use of inappropriate language or topics, chatbots are often censored using a 'blacklist' that prevents engagement with certain words or questions. This can result in problems of equity, inclusion and denial of service.

HR chatbots can be trained to use a specific database of employee language phrases to guide their conversation. However, they fall short in processing and mimicking written and spoken language which does not belong to the dominant group. Speech recognition from all of the major technology producers shows a significantly higher error rate with people who are black, misunderstanding between 25 and 45% of words spoken.[8]

3.4. Video interviews

The technology allows employers to interview job applicants on camera, using AI to rate videos of each candidate according to verbal and nonverbal cues. But the software reflects the previous preferences of hiring managers. So if more white males with generally homogeneous mannerisms have been hired in the past, algorithms will be trained to favourably rate predominantly fair-skinned, male candidates while penalising women and ethnic minorities who do not exhibit the same verbal and nonverbal cues.

Technology companies have already begun to make strong claims about the efficacy of facial movements to process emotion, and these are assumed to be cross cultural. However, the evidence base is limited in its reliability, lack of specificity and limitations to the generalisability. The technology is trained to address 'basic' emotions (e.g. anger, happiness, sadness), but is poorly designed for the more subtle, variable emotional expressions that may be conveyed in a job interview (e.g. confidence, enthusiasm), and almost completely ignore cultural differences in emotional expression. In addition, skin colour and face shape significantly alter results, with black profiles associated with anger or contempt, while Asian faces are perceived as blinking repeatedly (associated with nervousness or deceit). In the first case, the technology reproduces human bias, requiring that black professionals must amplify positive emotions to receive parity in their workplace performance evaluations. The second is based in face-shape perception by the technology. But skin colour can even affect whether an AI interview recognises a person is present, and whether it begins or continues the interview.

3.5. Assessments

Pre-employment assessments are used to measure aptitude, skills and personality traits to differentiate potential top performers from other applicants. Predictive assessment tools include 'off-the-shelf' assessments for a variety of job functions (like customer service, sales and project management) and competencies (like 'problem solving' and 'interpersonal skills'). Generic assessments are completed during the online application process, and are automatically scored 'with the help of machine learning' to predict generic job performance. Custom-built assessments use the employer's workforce and performance data to predict how new applicants may compare to current 'successful' employees.

Personality-related tools infer candidates' personality traits through a self-assessment survey, and score candidates on personal attributes and predicted alignment with an employer's desired traits. They cross-reference that information with the employer's own performance indicators for those employees (like employee reviews, promotions, or sales numbers) to identify the personality traits that most differentiate a company's high performers from its low performers. These produce a personality 'fit' but are not necessarily directly responsible for success in the related role, and could even be entirely circumstantial.

Game-based assessments and interactive activities may be used to infer these directly, rather than through surveys. 'Neuroscience' web and mobile games are used to measure cognitive, social and emotional traits of candidates, such as processing speed, memory and perseverance. Candidates may be assessed on the basis of their impulsivity, attention span and ability to learn from mistakes, even if they believe that they are assessed on the game results. **These assessments focus on selecting candidates that reflect current 'top performers' in the workplace, although there are no objective bases on which to identify 'top performers' that are free of bias.** Pre-employment tests measuring cognitive ability and personality (used for 'cultural fit' assessments) have long been suspected to be inherently discriminatory9 against racial and ethnic minorities as well as people with disabilities. Vendors of these products can take steps to mitigate observed disparities in their models, including using statistical techniques to remove obvious demographic biases when evaluating behavioural traits. However, as these models are not available for external auditing, confidence in their effectiveness is low.

3.6. Rank-ordering or scoring

Similarly, qualified candidates may be scored or ranked differently according to distinctions which are not particularly relevant to the selection stage. However, not enough is known about how rank-ordering affects human recruiters' decisions at this final stage. At the individual applicant level, there may be little difference between one applicant and another, but across a company or sector, it can affect final

[8] Koenecke, A., A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J.R. Rickford, D. Jurafsky, S. Goel. 2020. 'Racial disparities in automated speech recognition'. Proceedings of the National Academy of Sciences. April 2020, 117 (14) 7684-7689; DOI: 10.1073/pnas.1915768117; Metz, C. 2020. 'There Is a Racial Divide in Speech-Recognition Systems, Researchers Say'. New York Times, 23 March. Available at: https://www.nytimes.com/2020/03/23/technology/speech-recognition-bias-apple-amazon-google.html.

[9] United States Department of Labor. 1999. *Testing and Assessment: An Employer's Guide to Good Practices.* U.S. Department of Labor, Employment and Training Administration, Office of Policy and Research.

stage candidates repeatedly and persistently. Recruiters can be influenced by 'automation bias', which is the undue weight given to computerised data which is presented as precise and objective. This is particularly the case when recruiters are not confident about the impact of their own biases and increase trust in technical data which appears neutral.

3.7. Talent management and salary prediction

Like in recruitment, AI offers to give organisations more accurate and efficient predictions of a candidate's work-related behaviours and performance potential while in employment. These are used to increase internal mobility and internal candidate identification, and to support executive and senior management recruiting. Structural racism can be reflected in the value attributed to other biased outcomes, including industry recognition, board memberships, bonuses and awards. AI tools also offer the capacity to make compensation fairer, closing pay gaps between employees with similar education, experience and certifications. Predictive hiring tools are used to forecast a candidate's salary requirements, but also to assess pay at higher levels or after set periods. **These are affected by the same issues which affect the assessment of these characteristics for hiring, valuing some skills, awards and experiences over others in ways that reproduce structural racism and lead to limitations in career progression – the ethnic and racial glass ceiling.** Natural-language processing tools which help with employees' sentiment detection and react quickly to retain and engage employees who might leave are affected by the same issues as chatbots, assessing some language forms as being more valuable than others in a particular context.

3.8. Performance evaluation and disciplinary responses

Digital technologies gathering data from existing employees to augment recruitment with 'gold standard' benchmarks also create benchmarks for expected performance of current staff. This data can then be used for performance evaluation. While there is scope to support the career progression and development of employees – identifying gaps in understanding and compiling evidence for promotion – employees who fall below this 'gold standard' benchmark are flagged as being less productive. In many cases, this occurs in circumstances where their performance was not judged as problematic prior to the introduction of the new benchmark. This can lead to disciplinary processes and barriers to progression which were unforeseen, either by the employer or employee. Benchmarks for performance should take account of the distinction between 'gold standard' or 'ideal' characteristics, and acceptable standards for ongoing performance.

4. REMEDIES

In such a new and evolving field, identifying examples of best practice which encompasses all elements of the deployment of AI in HR would be limiting. However, some of the best learning can come from recognising the strong components of some tools and learning from the failure of others. Adoption of reflective practice is crucial here, and will depend heavily on a sense of capacity and confidence on the part of HR professionals. This section shows how existing successful bias mitigation techniques work, and how they can be used together to form effective mitigation strategies.

The European Commission High Level Group on AI has highlighted the following broad guidelines for ethical use of AI: Human agency and oversight; Technical robustness and safety; Privacy and Data governance; Transparency; Diversity, non-discrimination and fairness; Societal and environmental well-being; and Accountability. Any responsible use of intelligent technologies would be advised to identify steps to take which bring practice in line with these guidelines. As such, we will break down practical steps you can take which comply with these principles.

Task	Tech response	Risk	Solution
Online targeting	Classification and eligibility software (automation)	Targeting comes from characteristics of current workforce (data provenance), reliant on social norms which often reproduce cultural and structural biases, and eliminates more than includes	Seek data points outside existing organisational characteristics Vet the characteristics determined by the automated system Don't eliminate 'sensitive' points, but rather ensure they are not deciding characteristics, so that you can measure the diversity in applicants which are rejected also
Screening CVs	Classification and eligibility software (automation)	Reliant on social norms which often reproduce cultural and structural biases, eliminates more than includes	As above.
Setting up interviews	Chatbots correspondence, automated emails (automation, mainly)	Consent/ethics, may require biometrics for confirmation	Ask clearly for consent Clearly state that AI is communicating, not a person Offer override if someone needs to talk to a person
'Relationship building' with candidates	Some automation, more likely some emotional AI/deep learning depending on quality of chatbots	Consent, ethics of whether or not they know it is AI not human, validation measures may require biometrics, reliant on social norms which often reproduce cultural and structural biases	Clearly state that AI is communicating, not a person Provide opt-out alternatives for biometrics Address structural bias issues Include more potential candidates from underrepresented populations in criteria development
Tracking communication between candidate and employer	Chatbots, automated email responses	Don't get feedback, can't validate what the candidate says, some requirements for biometrics (associated ethical issues)	Clearly state that AI is communicating, not a person Provide opt-out alternatives for biometrics Build in regular human review
Video interview screening	Deep learning, emotional AI, video analysis	Unreliability of emotional AI, discriminatory psychometrics often the basis, expensive and labour intensive to maintain, requires elaborate equipment, requires huge data input for training, issues of privacy and consent, data can be personal or biometric	Provide opt-out alternatives for biometrics Provide structured, consistent interviews given identically Use an interview scorecard that grades candidates' responses to each question on a predetermined scale. Address structural bias issues, analysis and understanding of the problems of certain psychometrics – find counterbalance options Include more potential candidates from underrepresented populations in criteria development Don't use platforms with facial or voice inflection analysis tools that have been questioned or have not been vetted Roll out new interview tools and processes consistently, and once in use, ask candidates the same questions in the same ways, and evaluate them by comparing their answers horizontally for greater consistency

Screening of phone interviews	Deep learning, emotional AI	Unreliability of emotional AI, discriminatory psychometrics often the basis, expensive and labour intensive to maintain, requires elaborate equipment, requires huge data input for training, issues of privacy and consent, data can be personal or biometric	✓ Provide opt-out alternatives for biometrics ✓ Address structural bias issues, analysis and understanding of the problems of certain psychometrics – find counterbalance options ✓ Include more potential candidates from underrepresented populations in criteria development
Career progression	Classification and eligibility software (automation), deep learning	No-opt out, risks of penalising against high performers, collection of large scale data points – gleaning unnecessary data?	✓ Provide option of alternative route to progression (even if it has to be a special application) ✓ Undertake assessment of potential biases in-built in progression criteria and benchmarking
Performance evaluation	Classification and eligibility software (automation), deep learning	Benchmarks used to select candidates are used inappropriately to monitor and discipline employees without accounting for workplace dynamics and normal variation in practice	✓ Distinguish between gold standard for recruitment and a benchmark for regular performance, and ensure that this is differentiated in the commands and criteria used to build the algorithm for staff retention/progression

4.1. Benchmark to address bias

Successful algorithms are built upon the data points they are fed. To have a competent algorithm for recruitment, organisations must collect a large amount of candidate data points to provide a comprehensive view of a successful applicant. This can be done through candidate assessments. Developers and clients need to establish objective measures of competency and ‘fitness’ for a job role (performance benchmarks). Machine learning algorithms are trained to predict the likelihood of success in that job based on data from the interview. The model is tested to ensure it is valid in its analysis, based on the assessment answers. Test for bias in the output data before implementation begins in order to ensure no adverse impact against protected groups. Algorithmic bias describes systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one group of users over others. If any are found, you look in the data points for any factors that could be contributing to the bias and remove them from the model before retraining and retesting the algorithm. In this process, you are seeking to eliminate those algorithms that reflect “systematic and unfair” discrimination.

One technical solution offered by vendors is the Candidate Masking feature which hides evidence of personal characteristics from decision makers. With this solution, companies seek to prevent unconscious bias from impacting candidates at critical human-managed selection stages. But this solution does not address the structural and institutional racism which repeatedly limits the applicant pool and excludes candidates on the basis of characteristics – such as names, hobbies, schools, vocabulary – which act as a proxy for race, gender, etc.

4.2. Addressing the imbalance of information and capacity

Co-production is increasingly recognised as best practice in both the private and public sectors. Because homogeneity is often part of the issue HR teams are trying to address, this process should involve those who are potential candidates – in particular those from underrepresented populations within the workforce. Once again, consideration of the position of the user (in this case the user being the candidate) should be incorporated throughout. The idea of co-production in complex design systems and in corporate processes might seem impractical, however there are myriad ways in which affected groups could be involved without jeopardising corporate privacy and trade secrets. Developing user-vetted lists of desired characteristics and generating discussion within corporate teams about why these characteristics are deemed desirable are two concrete and relatively straightforward examples of this.

Commissioning staff who have appropriate capacity in the area of AI will be more confident in developing clear and explicit briefs which identify 'no go' areas in terms of personal data collection. It is important that employers seek to ask the question of designers – why do you need to collect this data. When tendering for off the shelf or ready-made tools, companies should ensure that they are looking for proof of application of ethical criteria, or incorporation of the seven guidelines on ethical AI outlined by the European Commission White Paper. The collection of personal data should be the minimum amount to achieve the desired outcome. This can be ensured when you commission tools by having a clear objective, and seek to achieve that objective through the collection of only clearly relevant data. Adopting these procedures is likely to futureproof businesses for the incoming Digital Services Act, which is likely to have a significant impact on the transparency of data collection and storage, as well as increasing liability for intermediary service providers. In order for practitioners to have the confidence to push back on a predictive analytics tool, a culture of openness must exist. A workplace culture must be cultivated in which HR practitioners are encouraged to challenge decisions – whether that be the decisions of other practitioners, their supervisors, or a predictive analytics tool.

4.3. Informed consent and opt-out

GDPR offers some protections in the right to access substantive information on the employment of automated decision making and data sorting, however this must often be considered in direct competition with the rights of the designers to maintain trade secrets and prevent copyright infringement. Best practice should have clear and practicable processes to address right to information requests. You should ensure that you are clear in the process of commissioning algorithmic design about what access you will have to the 'black box' element of more complex tools.

4.4. Increasing transparency and accountability

Core to ethical deployment of AI is the traceability and accountability of the processes. There is a risk that without the ability to walk through the steps that the algorithm takes, potential burden of proof placed on parties experiencing discrimination could become too onerous. Legal, technical and intellectual property barriers to identifying the how and why of algorithmic decision making tools – particularly tools which evolve at a rapid pace – could obstruct attempted discrimination claims and make challenging discriminatory practices more hidden and difficult to contest. Industry must seek to employ both legal and ethical frameworks, and to develop meaningful, practical ways to deploy new technologies to ensure that these principles are adhered to and embodied throughout the commissioning, design and implementation of intelligent systems. Employers who wish to stay abreast of best practice should seek to go further than mere compliance with a largely embryonic legal framework, and instead you should follow the lead of the European Commission's High Level Expert Group[10] in seeking a pro-actively ethical approach. This is particularly important for those of you who cite a desire to increase fairness in hiring practices as a key reason for the adoption of these tools.

4.5. Review aims for AI-assisted HR

The features of predictive hiring tools are designed to enhance your recruitment capabilities, not just in making your selections more efficient, but actually undertaking analyses which humans cannot do, and applying them to make new selection criteria. The problems we have laid out in this toolkit with understanding and changing the outcomes of those selections might lead employers to resist the temptation to adopt predictive tools. There is certainly a balance to be sought between looking for candidates that reflect successful employees in the sector, and ensuring that your organisation is not responsible for reproducing racial bias in the workplace, and for unlawful discrimination. Your aims might include using AI to increase workplace diversity, and many hiring vendors claim to be able to do so, either by avoiding discrimination against applicants in protected categories, or by proactively diversifying the applicant pool and selection process. Technical fixes often fall short of expected outcomes. This is why the process has to be led by the consumer, supported by reflective practice, organisational readiness and co-production in your organisation.

[10] The European Commission's High Level Expert Group on AI aims to support the implementation of the EU's Strategy on Artificial Intelligence. In 2018, the High Level Group produced Ethics Guidelines on Artificial Intelligence, which outlined seven key requirements for trustworthy AI. See: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Checklist

This checklist provides Human Resource Management commissioners of AI with a clear list of tasks to build confidence in developing a brief for AI driven recruitment and retention tools.

○ 1. **Ensure your brief includes explicit objectives of fairness and accountability**
Make it clear that fairness and accountability are important to your firm, and that you intend to make robust evaluation of the impact of these tools a priority in choosing a provider. With specific reference to the needs of your organisation, outline the equality and diversity policies and talk through how these should be upheld by the design of the automation and prediction tools. Hiring firms which demonstrate a track record of fair employment practices, and with a heterogeneous workforce is one way to uphold these principles. This should be applied both in bespoke and off the shelf tools.

○ 2. **Understand how the algorithm has been trained**
Algorithms can reproduce biases from the data sets which train them. Ask questions about what data sets are used to train the tool. Are these data sets from a diverse population? Are they sector specific? Is it an international data set, or restricted to one jurisdiction? The better understanding you have of the training data, the more confident you can feel about it.

○ 3. **Understand the data points used to complete the tasks**
Automated and predictive tools operate using simple commands (if:then). Many tools employ measures from psychometric tools or mimicking human cognition. Some of the categories and measures used to develop AI have been shown to have racialised, gendered or culturally specific outcomes. Ensure you understand the commands which drive your tool, and make suggestions and adjustments where possible and appropriate.

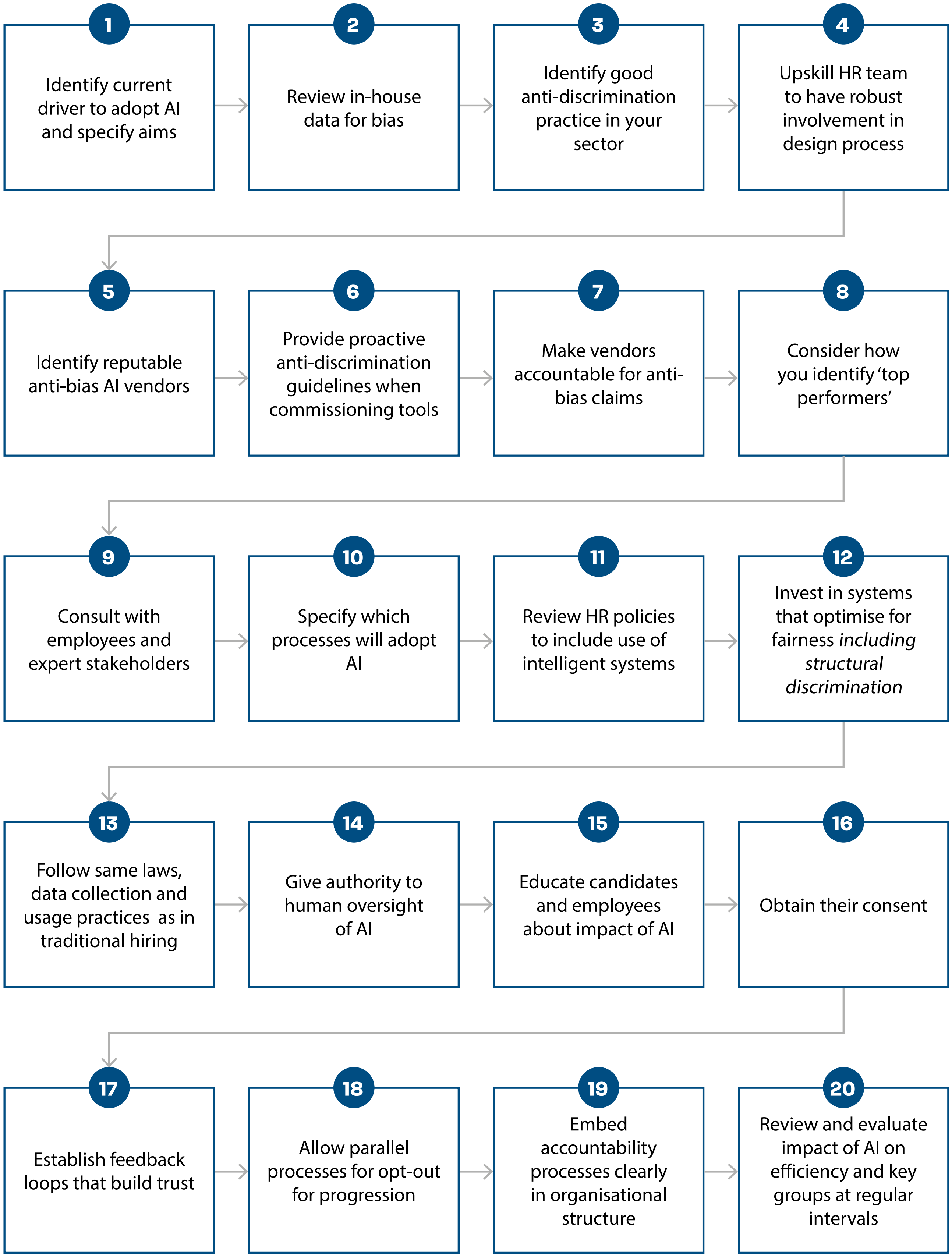
○ 4. **Negotiate influence over the training data and commands**
If you are commissioning a bespoke tool, negotiate a level of input into what training data is used, and take an active role in identifying what commands or criteria will be used in its development. More complex, deep learning machines will have minimal input prior to training, so in these cases, ensure you understand how you and your team can review and evaluate outcomes on a regular basis.

○ 5. **Ensure that there is a robust and inclusive consent process, including opt-out**
Go beyond GDPR and privacy regulations and recognise that for many job candidates who really want to work for your organisation, opting out of these processes won't be much of a 'choice'. Ensure that candidates have access to a clear, concise information page with a range of options to opt-out of deep learning processes where possible. Incentivise the use of AI (through ease of use and enhanced user experience), but offer human-based alternatives for those with genuine concerns.

○ 6. **Ensure that there are clear and accessible pathways for human override**
Candidates should have access to as much information as possible about how decisions on their application were made, and a right to appeal this decision. Without having access to the (sometimes thousands) of factors which go into a deep learning screening process, ensuring that there are practical human override options will be key. Policies and practice which respect human override will be necessary to ensure these options are utilised where appropriate.

○ 7. **Build in an actionable review of predictions and outcomes**
Many off the shelf tools will have limitations to how much information the vendors are willing to share about the algorithm and its processes. Without this transparency, it will be necessary to identify a way to review the predictions and decisions of the tool. Negotiate with the vendor to ensure that rejected applications are accessible for human review at random, acting as a quality control system and means of identifying problems before they escalate.

Model roadmap



1. Identify current driver to adopt AI and specify aims
2. Review in-house data for bias
3. Identify good anti-discrimination practice in your sector
4. Upskill HR team to have robust involvement in design process
5. Identify reputable anti-bias AI vendors
6. Provide proactive anti-discrimination guidelines when commissioning tools
7. Make vendors accountable for anti-bias claims
8. Consider how you identify 'top performers'
9. Consult with employees and expert stakeholders
10. Specify which processes will adopt AI
11. Review HR policies to include use of intelligent systems
12. Invest in systems that optimise for fairness *including structural discrimination*
13. Follow same laws, data collection and usage practices as in traditional hiring
14. Give authority to human oversight of AI
15. Educate candidates and employees about impact of AI
16. Obtain their consent
17. Establish feedback loops that build trust
18. Allow parallel processes for opt-out for progression
19. Embed accountability processes clearly in organisational structure
20. Review and evaluate impact of AI on efficiency and key groups at regular intervals

Glossary / Key concepts

Algorithm: A set of precise instructions that describe how to process information, typically in order to perform a calculation or solve a problem. Algorithms have to be described in programming language to be executed on computers.

Algorithmic bias: The systematic, repeatable behaviour of an algorithm that leads to the unfair treatment of a certain group.

Artificial Intelligence (AI): An area of computer science that aims to replicate human intelligence abilities in computers. Definitions focus either on achieving human performance in complex tasks, or on mimicking the ways in which these tasks are performed by humans. In a commercial context, AI currently refers mainly to systems that use machine learning for pattern detection, prediction, human-machine dialog, and robotic control.

Attribute: A variable used as part of the description of a data sample or classifier, for example a specific pixel in a camera image, or the gender column in a spreadsheet describing employees.

Deep learning: A neural network with many layers of nodes, each of which is capable of detecting patterns at different levels of abstraction from the previous one. Deep neural networks have been used to achieve or surpass human performance in very complex tasks. They typically require very large amounts of training data. The models learnt by deep neural networks are very hard to inspect, interpret and explain; they currently remain largely opaque.

Direct discrimination: The process of consciously and explicitly using group membership when making decisions about an individual. In legal definitions, direct discrimination occurs where one person is treated less favourably than another is, has been or would be treated in a comparable situation on the ground of a protected characteristic.

Discrimination: The process of making distinctions in the treatment of different individuals based on their actual or perceived membership to a group or social category.

Fairness: Impartial and just treatment without favouritism or discrimination in the most general sense. A complex concept that is associated, among other things, with notions of: equitable, non-discriminatory treatment in legal and administrative processes; fair distribution of wealth and other societal benefits based on concepts like social justice, solidarity and compassion; and appropriateness of treatment in interpersonal interaction, linked to respect and universal rights.

Machine Learning (ML): The science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions. Instead of requiring explicit programming of this model, ML algorithms identify patterns in data to develop a model that can be used to reproduce or predict the behaviour of the system they are trying to learn about. When provided with sufficient data, a machine learning algorithm can learn to make predictions or solve problems, such as identifying objects in pictures or winning at particular games.

Model (machine learning): A mathematical representation of a real-world process. This may be a 'hypothesis' regarding a phenomenon described by data, that ideally provides a concise explanation of complex observations by identifying generalisable patterns and ignoring irrelevant variations.

Protected characteristics: Attributes of individuals explicitly protected by anti-discrimination law. These can include, depending on jurisdiction, age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, and sex.

Race: The socially constructed classification of humans into groups based on physical traits (such as skin colour), ancestry, religion, genetics or social relations, or the relations between them.

Racism: The prejudice, discrimination or antagonism directed toward someone of a different race, based on the belief that one's own race is superior. Racism, as an ideology, exists in a society at both the individual and the institutional level. Consequently, the systemic nature of racism, as well as who holds the power to perpetuate it, is becoming more popular in mainstream discourses of the term.

Structural discrimination: Refers to a range of laws, policies, rules, attitudes, and behaviours in institutions and society which cause barriers and prevent equal access to rights and opportunities for minority groups. Structural discrimination is often aligned with privilege and disadvantage aligned with societal norms, power and dominance related to race, gender, religion, sexuality, and other social, economic and cultural power relations.

Further reading

Bargain, Ch., Beaurepaire, M. & Prud'homme, D. 2019. *Recruter avec des Algorithmes: Usages, Opportunités et Risques*. AFMD.

Bogen, M., & Rieke, A. 2018. *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*. Upturn.

Brione, P. 2020. *My Boss the Algorithm: an Ethical Look at Algorithms in the Workplace.* Acas Research Paper.

Council of Europe. 2018. *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making.* Strasbourg: Council of Europe.

Courtland, R. 2018. *'Bias Detectives: the Researchers Striving to Make Algorithms Fair'*. *Nature*, 20 June.

Engler, A. 2020. *The European Commission Considers New Regulations and Enforcement for "High-risk" AI. Brookings Institute,* 26 February.

Equality and Human Rights Commission. *Algorithms and Artificial Intelligence Reading List*.

Eubanks, V. 2018. *Automating Inequality.* St Martin's Press.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A.C. and Srikumar, M. 2020. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Berkman Klein Center for Internet & Society Research Publication Series. Harvard University.

Jefferson, B. 2020. *Digitize and Punish: Racial Criminalization in the Digital Age.* University of Minnesota Press.

Mateescu, A., & Nguyen, A. 2019. *Algorithmic Management in the Workplace*. Data&Society.

Moore, P. V. 2018. *The Threat of Physical and Psychosocial Violence and Harassment in Digitalized Work*. International Labour Organization.

Nesta. 2019. *Decision-making in the Age of the Algorithm: Three Key Principles to Help Public Sector Organisations Make the Most of Artificial Intelligence Tools*.

New Zealand Government. 2020. *Algorithm Charter for Aotearoa New Zealand*.

O'Neill, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin.

Rovatsos, M., Mittelstadt, B., & Koene, A. 2019. *Landscape Summary: Bias in Algorithmic Decision-Making*. Centre for Data Ethics and Innovation.

Ruha, B. 2019. *Race after Technology. Abolitionist Tools for the New Jim Code.* Polity.

Williams, P. and Kind, E. 2019. *Data-driven Policing: the Hardwiring of Discriminatory Policing Practices across Europe.* European Network Against Racism (ENAR).

World Economic Forum. 2018. *How to Prevent Discriminatory Outcomes in Machine Learning*. Cologny, Switzerland: World Economic Forum Global Future Council on Human Rights 2016-2018.

Notes

#20112 • crossmark.be

european network against racism aisbl

Tel: +32-2-229.35.70 • E-mail: info@enar-eu.org
/ENAREurope • @ENAREurope

The European Network Against Racism (ENAR) stands against racism and discrimination and advocates equality, solidarity and well-being for all in Europe. We connect local and national anti-racism NGOs throughout Europe and act as an interface between our member organisations and the European institutions. We voice the concerns of ethnic and religious minorities in European and national policy debates.

Visit ENAR's website: www.enar-eu.org

ENAR foundation

This report was supported by ENAR Foundation. You can support its work towards achieving a racism-free Europe by donating online: www.enarfoundation.eu