Biased by Design? Unlocking AI's Hidden Risks

13 November 2024

The integration of AI systems into the workplace has brought about significant advancements in increasing efficiency across HR systems and lowering costs for employers. However, there is a growing concern around how biased data can make its way into AI systems, with potentially harmful and discriminatory results.

In this article in our **AI series**, we explore some of the key challenges surrounding AI bias, the reasons why they occur and how organisations can address them.

What is AI Bias?

AI bias refers to unfair outcomes or systematic errors in artificial intelligence systems that can lead to unintended and discriminatory effects on groups based on their characteristics such as race, age, disability or gender. AI bias can stem from various different sources, including the data used to train AI models, the design of algorithms themselves and human bias.

Unintended bias in AI algorithms could arise if AI systems are being used for hiring, promotions, performance evaluations, or terminations in the workplace. AI systems are typically trained using existing datasets. Often the internal workings of the AI system are invisible to the user and it can be particularly challenging to unpick why a decision was reached by AI and what weight was given to certain factors. Bias can creep in when the data the AI learns from or the rules it follows contain human biases.

By way of example, if an AI hiring tool is trained predominantly with male resumes, it will likely favour male candidates, inadvertently excluding or downgrading female applications. There is an obvious risk of past discrimination and historical bias being embedded and exacerbated in any algorithm. This bias can then lead to work related discrimination claims.

Risk of Unlawful Work Related Discrimination

In Ireland, the Employment Equality Acts prohibit discrimination in the workplace, whether direct or indirect, on the basis of the nine protected grounds – gender, marital status, family status, age, disability, sexual orientation, race, religion and membership of the traveller community. An employee or job applicant who feels that they have been discriminated against unlawfully on any of the prohibited grounds can make a claim. This can include a situation where an individual believes that an employer's use of an AI tool in an employment related decision resulted in a discriminatory outcome.

If an individual brings a discrimination claim, an employer must be able to show the reason it took a particular action, such as a decision not to recruit or promote. If it cannot explain the reason for the decision (e.g. because it is unclear how the AI arrived at certain outcomes), an adverse inference may be drawn that the reason was a particular protected characteristic and the individual may succeed in their claim.

There is a real potential for certain workplace use cases of AI to put groups who share a particular protected characteristic at a disadvantage compared to others by reason of that characteristic. For example, an automated timed assessment in a recruitment selection process to assess problem solving skills could disadvantage a candidate who was neuro diverse, such as someone with ADHD or dyslexia, who may require take longer to respond due to differences in information processing, potentially scoring lower than peers with the same skill level. This is indirect discrimination.

In general, direct discrimination can't be objectively justified. However, in indirect discrimination cases, an objective justification defence may apply if it can be shown that the deployment of the AI system was a proportionate means of achieving a legitimate aim. The proportionality element of the defence can be challenging to establish in practice as it involves a careful balancing exercise to determine whether the discriminatory impact is outweighed by the needs of the user of the system.

Employers should also be particularly mindful of the obligation to provide reasonable accommodation to disabled employees and applicants and should assess whether usage of AI in a process may be disadvantageous to certain disabilities and what adjustments may be required to alleviate the disadvantage.

In Ireland, compensatory awards for the effects of discrimination can be up to 2 years pay or up to €13,000 for someone who is not an employee, such as a job applicant. There is also potential significant reputational risk for businesses as a consequence of AI bias.

In a recent high profile UK case, an online food delivery platform agreed a financial settlement with an employee following allegations that its facial recognition software used to access their work app was "racially discriminatory". The employee experienced difficulties with the platform's verification checks, which use facial detection and recognition software. The AI tool prevented the employee from accessing the app to secure work due to "continued mismatches" in the photos of his face he had taken for the purpose of accessing the platform.

Another example of AI bias creeping into HR systems resulted in a US company agreeing a significant settlement of hundreds of thousands of dollars. A job applicant discovered that the company's AI screening platform used during the hiring stage was discriminating against older applicants – rejecting applications from women over 55 and men over 60. The applicant in question had re-applied using a fake younger age and his application was accepted. This automated discrimination led to the rejection of over 200 qualified applicants – with AI deciding whether a candidate is a good match or falls short.

Human Input

These examples serve as important precedent on the legal and reputational implications of using AI to automate hiring and other employment functions. The recently refreshed Irish National Strategy for AI 2024 emphasises the importance of better awareness and transparency among businesses and users of

how trustworthy AI functions and how individual rights are protected. Organisations should not overlook the importance of AI literacy and keeping the human in the loop at all stages of the design and implementation of AI tools by running checks on the AI's output and ensuring that certain uses of AI in employment settings are regulated. Ultimately, the AI and the human must work together. Contrary to the notion of machines taking over from employees' positions, the reality is that humans play a significant role in the regulation of AI systems.

The EU AI Act

The EU AI Act includes specific provisions for bias detection, requiring that high-risk AI systems undergo rigorous testing and validation before their deployment in the EU marketplace. Employers using AI tools in the field of recruitment or decision-making in the workplace will need to engage with new obligations.

In keeping with the human-centric approach to AI and to ensure the protection of fundamental rights – such as non-discrimination – one such obligation that will be introduced under the EU's AI Act is a new requirement on businesses to carry out an assessment of the impact of fundamental rights that the use of an AI system may produce before using such a system ("**FRIA**"), where applicable (eg, if a public body, providing a public service or operating in banking and insurance). Within this FRIA, employers are asked to identify the specific risks of the AI system to the rights of individuals or groups of individuals likely to be affected and identify appropriate measures if these risks materialise. FRIAs must be performed before putting the high-risk AI system into use and need to be continually reviewed and updated throughout the lifecycle of a high-risk AI system.

Employers, as "deployers" of an AI system, will also be subject to a new right for individuals to obtain a clear and meaningful explanation of the role of any high-risk AI system in decision-making they have been subject to, under the EU AI Act. It remains to be seen how this right may be used in practice. It is possible that it may be used by employees to put pressure on employers to explain algorithmic decision-making in the workplace and possibly to support potential employment related discrimination claims – one to watch in the future. That being said, employers already have an obligation under the GDPR to inform employees about their use of any AI systems used to make automated decisions about them so this new right may not subject employer's to anything more onerous than what they already should be doing under their GDPR obligations.

For further information on the EU AI Act and the new obligations for employers, please see our previous insights article here.

Conclusion

To reduce the risk of bias and discrimination claims, there are a number of practical steps employers can take. In the design and implementation phases of AI systems, employers should conduct an assessment of any potential bias risk and/or adverse impacts that the proposed usage may have on certain protected groups, any objective justifications for differential treatment and remedial steps that may be taken. Clear policies on the use of AI systems in the workplace should be introduced and those involved in implementing AI tools in the workplace should be appropriately trained on equality laws and transparency obligations.

Matheson's Employment, Pensions and Benefits Group is available to guide you through the complexities of navigating the ever increasing use of AI in the workplace and new legislation being introduced in this area, so please do reach to our team or your usual Matheson contact.

Our Digital Economy Group have published a comprehensive AI Guide for Businesses, which provides an overview of the AI Act, and will help you to understand the scope of your new obligations. To get your copies of the Matheson AI Guide for Businesses, please email: AIGUIDE@matheson.com

This article was co-authored by Employment, Pensions and Benefits partner, Alice Duffy, associate, Ciara Taggart and trainee solicitor, Niall Quinlan.