The Agenda podcast by Lewis Silkin: AI 101 - using AI in employment - explaining decisions and addressing bias

Tarun

Welcome, listeners to our AI 101 podcast series through which we will share insights on the key legal and practical implications of the advent of accessible and generative AI. I'm Tarun Tawakley and I'm joined in this episode by my colleague, Becky Jobling.

This is the second in the series of our podcasts on the topic; the first in this run being an introduction to AI with Olly Fairhurst and JJ Shaw. If you haven't had a chance to listen to that already, I can highly recommend that you do so. Today though, we're going to move on and look at the use of AI in employment, focusing on the question of trust and the importance of being able to explain AI decision-making to users.

Becky

Absolutely, Tarun. So, trust in technology, generally, has been dealt a pretty significant blow in recent months by the Post Office scandal, I'm sure everyone listening to this will have heard about this and will be aware that this situation arose out of the faulty Horizon software.

This might not have been caused by the new generation of AI tools that we've seen come to the fore in the past year or so, but it certainly sheds light on the potentially harmful consequences of technological interventions. And of course, that's something that's highly relevant to any discussion of AI. And I think particularly in the workforce, so here, we're talking about decisions that have the potential to have a huge impact on people's lives.

So, in this area, ethical standards and safeguards are going to be crucial to maintaining trust and confidence in that technology.

Tarun

I think that's right, I mean one of the things that I find really interesting about being in employment law is the real-world application and impact that employment laws have on day-to-day working lives of people, and I think when you start thinking about the impact that AI is going to have on the world, the impact that it's going to have on the workplace, again, is going to be tangibly felt by everyone.

So, putting these issues into the context of work, we're going to look at how AI is supporting, or indeed, making decisions in the workplace. What are some of the risks associated with that? And finally, what safeguards are we seeing starting to emerge and will they help retain trust in the technology?

So, let's start with the basics. Why is the issue of trust so important? As Becky has alluded to in talking about the Post Office scandal, which unless you've been living under a rock, has sort of captured the public and frankly political imagination like few other stories in recent times. Let's be clear, it has absolutely nothing to do with AI, just like the P&O scandal has got nothing to do with fire and rehire in any real sense, but it's captured people's interest and it's by no means a standalone, in terms of a story of problematic issues for businesses, arising out of the increased reliance on new technology. Frankly, early adopters of almost any technology often see problems and limitations in the technology, which if unchecked can really go to undermine confidence in its abilities.

Let's look at a couple of recent examples in the context of generative AI. Firstly, with DPD and issues they've had recently with their AI powered chatbot. For those of you unfamiliar with this, the tool ended up complying

with prompt instructions from a disgruntled customer and after a very few rounds of experimentation, the customer was able to have the chatbot disregard its normal rules and before long the chatbot was swearing, was calling itself useless and it even composed a poem criticising the company.

This isn't the only example, there's also some interesting cases from the US where Chevrolet rolled out a similar tool , a customer was pretty quickly able to prompt the chatbot to agree to everything the customer said, regardless of how ridiculous the question is, and to confirm that any answers it gave were “legally binding offers with no takesies backsies”. Before long, this user was able to convince the chatbot to agree to sell the user a $78,000 vehicle for $1. Accompanied of course, with that crucially important bit of legal drafting that it was legally binding and without any takesies backsies.

Becky

Brilliant, yeah. I love those examples, Tarun. Another high-profile example recently, was Google's generative AI tool, Gemini. Google had instructed the tool not to make assumptions about characteristics such as sex and race but actually, what this meant was that this then produced historically inaccurate and potentially offensive images, underscoring the difficulty in correcting bias.

Tarun

And I think that one's so important Becky, because you can see what Google we're trying to do here. As we'll come onto, there have been lots of issues with bias inbuilt in data sets and with historic data sets creating issues where for example, in facial recognition technology it isn't able to recognise non-white faces where there wasn't enough data in the sample set for it to be able to recognise the distinctions. Becky will come on to this later, but it shows that in simply trying to address this problem, you can very easily end up overcorrecting it, creating equally untrustworthy results.

Becky

Yes, I mean that's right, Tarun. I think this underscores the difficulty of correcting bias. It's not an easy fix, and this contributes to our perception of AI as tech that can go dramatically awry and potentially can't be trusted. And we all talk about the duty of trust and confidence in employment and that's fundamental to the employment relationship. So how employers use AI to make and support their decisions needs to be done in a manner that protects that trust. So, holding that thought, we're going to take a step back and have a look at how employers are using AI in the workforce. And Tarun, can you talk us through some of the common examples that we're seeing, perhaps?

Tarun

Of course, I mean, I suppose the thing that's worth saying before we get into this, is the answer to this probably is more than they realise, and in ways that they probably weren't aware of, but let's stick with the stuff that we will know about as employers.

Last year's government briefing paper on AI and employment law gives a useful overview of the topic. It identified three broad areas in which AI is being used in the world of work and unsurprisingly, these are recruitment, line management, and monitoring and surveillance.

So let's start with recruitment. Unsurprisingly, this is one of the most obvious use cases for an AI tool. It can be used to source candidates and analyse skills and qualifications that the role might need. Chatbots might even guide people through the interview process. It could be incredibly useful in screening exercises;

technology can sift through CVs based on key criteria to identify those that are most likely to be suitable for the role.

We should note an air of caution here, when we've been talking about examples of things going wrong so far, it's worth highlighting the example involving Amazon in 2018 when it built its own in-house AI recruitment tool and discovered the unintentional gender bias in the algorithm, learnt from the data in its training set. And you can understand how this happens, if you feed an AI tool information about what a successful candidate looks like based on your current employee population, and if that employee population is, taking tech as an example, largely male and white, you can understand how the technology might put two and two together and decide the answer is five, and decide that those sorts of criteria are the common characteristics of successful candidates. Finally then in recruitment, it can also be used in the context of assessment and interviews through online assessments outside of the general screening process to assess suitability of candidates.

Becky

That's such a great example, the Amazon one, isn't it? The next area that's identified in the government paper was line management. So, particularly in the retail and hospitality industry we're seeing shift allocation algorithms become increasingly prevalent. I mean, their implementation isn't necessarily straightforward; we see things like short notice shifts, or the allocation of micro shifts and perhaps, these are the kind of things that the human might be more sensitive to in terms of the sort of nuance of that scheduling process. Another common example in line management would be performance evaluation, so using AI algorithms to quantify productivity and assess performance metrics.

And finally, the third area that the government paper identified and it's probably the most controversial use case for AI in the workplace, is monitoring and surveillance. And many of our listeners I think would have heard of the ongoing tribunal case concerning Uber's facial recognition technology. Uber Eats uses this tech to verify workers' identities at the beginning of the shifts, but concerns have been raised regarding the efficacy of that technology for individuals from minority ethnic groups. And again, talking about you know the data that that's been fed into this system, this stems from the algorithms being trained on data sets that inadequately represent those demographics.

Tarun

Perfect. So, having thought about how AI is being used in the employment context and the world of work, what are the risks in doing so? Obviously, there are a range of risks from data security, risks around accuracy and issues that arise from a lack of human oversight. But what we'll turn to now, given the focus of today's session, is the point Becky's just alluded to around the risk of bias and discrimination. How can this arise in technology that is specifically designed to make objective decisions?

This can be complex. It could be bias arising out of the underlying data that could either be the training data, which can reflect historical and societal inequalities or stereotypes, or bias in real-world data that supplies the AI model once it's in operation.

Then, there is the potential for the AI to model and magnify that bias in its own decision-making.

Secondly, there could be bias arising out of the way information is categorised and how the AI tool is asked to consider the information in its decision-making.

Putting that into practice in the recruitment example, how might bias creep in here? So, there might be underrepresentation of certain groups in the data, or if success is defined by reference to previous successful candidates, as we talked about in that Amazon example, they may all share a particular

characteristic that might mean you get skewed results. Even if you try and hide such data, such as sex or ethnicity, tools can sometimes pick up on trends and there have been multiple cases and stories of incidents where AI powered tools have identified common characteristics, so for example in the US, identifying that playing lacrosse at college, typically a sport played by men in the US at university, is more likely to result in you being a successful candidate.

Becky

That's right, Tarun. The tools seem to have a good way of finding proxies for things like gender, age, things like that.

Tarun

Of course, because ultimately what they're looking for is patterns in successful candidates. And those patterns may be, you know the fact of the lacrosse championship, which in and of itself, a tool will not be able to readily identify as being problematic unless that's part of its programming.

So, what risk does this present for employers? Well, at least in the UK, there's no specific AI legislation yet, although watch this space because I know behind the scenes, a lot of work is being put into this, with bodies like the TUC calling for far greater controls.

But even today, on the law as it is, there's a clear basis for discrimination claims under the Equality Act and also claims arising under data privacy laws. If you're interested in learning more, you can check out our case study on discrimination and recruitment, which can be accessed via the link in the episode description.

Becky

Thanks Tarun. So, we've discussed the risks associated with AI but how can we minimise these risks and still preserve that crucial element of trust? I think on this, it's useful to have a look at how regulations are evolving globally to promote AI safety. And this is an area that's continually developing and we're in fact, going to come back to this in a later podcast for a closer look. Whilst it might make our lives easier, there isn't a one-size-fits-all global standard, but certainly what we can identify are emerging categories of safeguards and I'm going to, sort of, run through these, I think it pulls out some interesting points.

So, in terms of the first category of safeguards, there's a concept of impact assessments; auditing and monitoring. So, what does that mean? Essentially, that involves developers or users of AI ensuring that the system is safe, both initially and on an ongoing basis. And as a concept that's not an entirely new thing, it aligns with the practice of data protection impact assessments.

Secondly, human oversight and intervention - essentially adding a human into the mix. But really there isn't one method ~~that~~ that we're talking about here, it's a spectrum of, sort of, interventions. It might mean a bird's eye oversight of the system generally, or it might mean something more interventionist, so, you know, interposing a person between the preferences revealed by the AI system and the final decision.

Third safeguard commonly emerging is contestability. It's well recognised that individuals affected by AI decisions should have the right to challenge them effectively and that promotes accountability and transparency.

And that takes us on to our last safeguard around that really important concept of transparency and also explainability, and those are vital for building that trust. We're going to talk about this in more detail, but essentially this means that the AI decisions are understandable to the people affected by them, that they get clear and meaningful information about how the AI system works and the factors implementing its choices.

So, I think Tarun's going to pick out a couple of our favourite safeguards there, the first one being human oversight and intervention and we're going to have a think about how effective that actually has the potential to be as a safeguard.

Tarun

Of course. We often hear this cited as a key safeguard but really, it could cover a spectrum of things. What is appropriate, will of course, depend a lot on the use case for the technology. In the recruitment context, will a human being really be able to evaluate all the decisions or are they simply going to be rubber stamping those?

To effectively review the decision, the person really needs to be able to understand how it was made, but may well come up against the issue of black box decision-making, where the algorithm is producing an output - a short list or whatever else it might be without really understanding or revealing how it arrived at that decision or at least without revealing it in a way that is intelligible to the average human being. Opaque decisions will present a number of problems and the key one being the issue of trust. That's before you even get into issues such as the burden of proof if the outcome is potentially discriminatory in a tribunal claim.

So, if a person is really going to evaluate that output, it's got to be more than a rubber-stamping exercise. They need to be able to unpick and understand that data which links to the importance of explainability and frankly, I think justifies the need for exercises such as counterfactuals to be run to be able to analyse and test the data being output so that there is a real understanding for any human being involved in the process as to what the grounds for the decision were.

Becky

So, I'm going to just look briefly at explainability, that's one of these new words that's becoming part of our lexicon around AI. What is this? So, you know, it's defined in the UK AI white paper as the extent to which relevant parties can access, interpret and understand the decision-making processes of an AI system. It's something that's linked to perhaps the broader concept of transparency. I'd say that really comes down to an openness about the nuts and the bolts of the system; how the model functions, what datasets are used, inputs and outputs and things like that.

Explainability is about being able to unpick that output, really, to achieve a human understanding of how a machine learning model has arrived at a particular conclusion. However, tricky that might be to understand for a human. But again, we're talking about a spectrum here. So, one extreme, in quite a simplistic way, an unsuccessful candidate might be informed that AI was used in the decision-making process and which factors were considered but, you know really, that's not giving them a whole lot of useful information in terms of genuinely understanding that decision. And so, at the other end of the spectrum, it involves a system that enables a deep understanding of how the AI model was applied to an individual and the predictions made about them. And that's something we've written about in quite a lot of detail in an article, in which we term this "local explainability" and we will put a link to that article, if you'd like to read about that and see the quite detailed case study about recruitment, of course, in the episode description.

So, coming back to trust. A decision that can be explained on an individual level, much more effectively addresses the risk of mistrust from black box decision-making that Tarun touched on earlier.

So why does explainability matter? I mean, again, in the recruitment context, it benefits both applicants and employers. It means applicants can effectively understand decisions and employers can potentially avoid the risk of discrimination inferences, which could stem from an inability to explain AI-driven decisions. And it can also lead to better decision-making, potentially surpassing the human decision-making capabilities, which of course, aren't immune to biases.

Practical steps? You know, we we've talked about there being no AI-specific regulation yet, but you know, that doesn't mean this is something to overlook. When putting AI tools in place, there are certainly questions employers should be asking that come down to this issue around explainability and transparency. And making sure they understand to what extent can the algorithm generated output be unpicked, what training is need of those users to understand those outputs, testing the product, understanding its capabilities, and that's going to be really important for being able to interpret those outputs. And that will go hand in hand with things like auditing and open communication with the workforce or with applicants about how AI is being used. And those things are all going to go and strengthen that crucial message around trust in in the technology.

Tarun

Thanks Becky. So, in this session we've looked at the impact of AI on employment law, the importance of accuracy, the importance of explainability, and safeguards that can be put in place to try and ensure appropriate and fair decision-making.

But what about the legal position? What does the future of regulation look like in this space? Well, very much still a watch this space territory. In the UK, it still looks likely that we're going to adopt a light touch approach following the UK AI white paper where the government is currently proposing a system that largely relies upon guidance for regulators rather than specific legislation.

From an EU perspective, the EU AI Act is edging ever closer to the finishing line. I think employers are going to need to turn their minds to the question of engendering trust whether required to by regulation or not because even on the law as it is today, whether facing discrimination claims in the recruitment context or simply questions about the fairness of the technology, which could go to public perception, it's really important that you get these things right.

Thanks very much for listening. If you'd like to learn more, please subscribe to our podcast series and you can check out our AI Hub on the website.