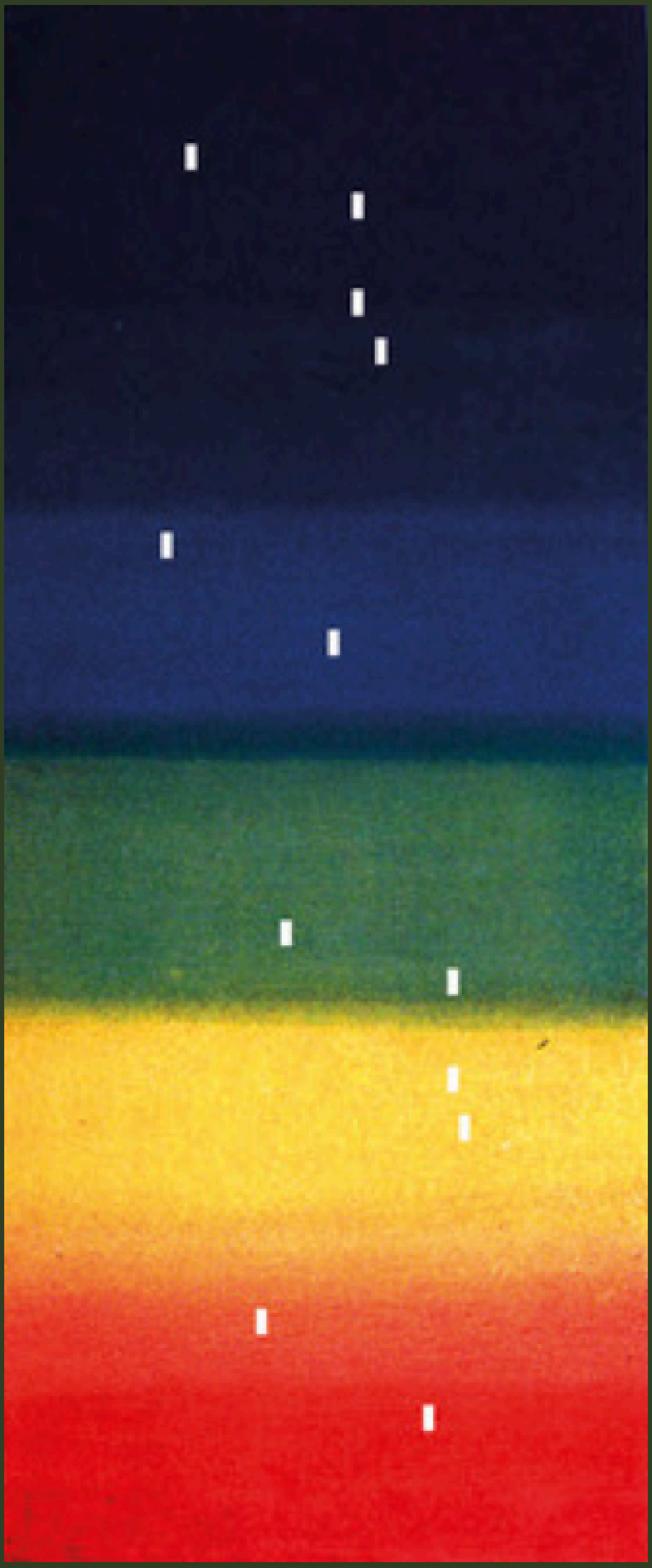
Jack&Jill



2026

A guide to bias in AI for recruitment

Contents

01
The impossible job.

You're in a bind.

Applications per hire have tripled since ChatGPT launched in 2022. Most of your roles get overloaded with AI-polished CVs that give you little chance of picking the great candidates from the bad.

"AI will solve it, you need to embrace it," says leadership.

You're open to AI, there's no other option. But you're uneasy about bias. You know the models regularly get it wrong, going off shallow signals and are only as good as the data they're given. So you check the outputs – three, four times. You lose time and your anxiety grows.

The resources to do it manually aren't there.

You're already stretched thin, and hiring another recruiter to handle the volume isn't in budget.

There are alternative paths. There have to be. Hiring with AI can be fair and effective. First, you need to understand where bias shows up, so you know what you're up against.

02
Bias is not a new problem.

Human bias is still here.

Bias in hiring isn't an AI problem. It's a human problem that now also manifests in AI.

Decades of field experiments tell a consistent story. In 2004, CVs with white-sounding names got 50% more callbacks than identical CVs with Black-sounding names, resulting in a gap worth eight years of experience. Remarkably, it hasn't got any better. A meta-analysis spanning 1989–2015 found no improvement over that period. Recent studies at major employers confirm discrimination persists today.

Gender, disability, age, physical appearance all show complex and intersectional patterns of bias, and they're real. The conditions under which hiring happens make it worse. Time pressure, high volumes, and imperfect information all contribute. Screening at scale amplifies these patterns.

This is the baseline. This is what "normal" human hiring looks like.

New AI tools, same old bias.

Early AI hiring tools learned directly from historical decisions and inherited the biases baked in. Amazon's CV screener downgraded any CV containing "women's" before the company scrapped it in 2018. HireVue's video platform faced complaints that it performed worse for deaf candidates and those with non-standard accents.

But general-purpose LLMs like ChatGPT and Gemini present a different problem. They're not trained on hiring data specifically, but they absorb biases from internet-scale text. Testing on over 500 job listings found LLMs favoured white-associated names 85% of the time and never favoured Black male names over white male names in any job category. Another study of 361,000 CVs found lower scores specifically for Black male candidates across GPT-3.5, GPT-4o, Gemini, Claude 3.5, and Llama 3.

Bias patterns vary significantly by model, job type, and candidate profile. Some less-biased models are also less effective at making good hiring decisions overall. You can't just measure bias in isolation. You have to assess decision quality at the same time.

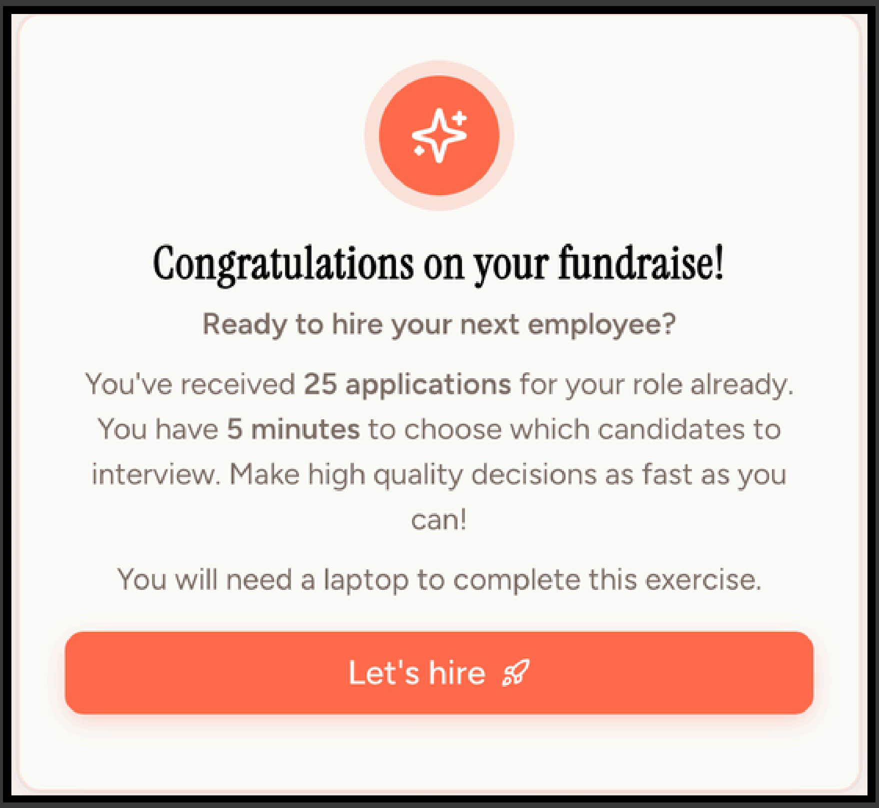
We investigated bias ourselves

Whilst the academic research is clear, it's mostly from controlled lab settings or historical hiring data. When we built our AI agent recruiter, Jill, back in July, we wanted to do our own research.

We built a game to recreate some of the pressures TAs face. Players had three and a half minutes to review 25 CVs and select their top candidates for a role. The results were uncomfortable. When given equally qualified candidates, humans selected the man 35% of the time, compared to just 25% for the woman.

To test AI models, we gave 13 LLMs the same set of CVs. They evaluated the same candidates differently 23% of the time when only demographic signals changed. In other words, they were remarkably inconsistent in how they judged candidates.

The game's still available to play to try for yourself.

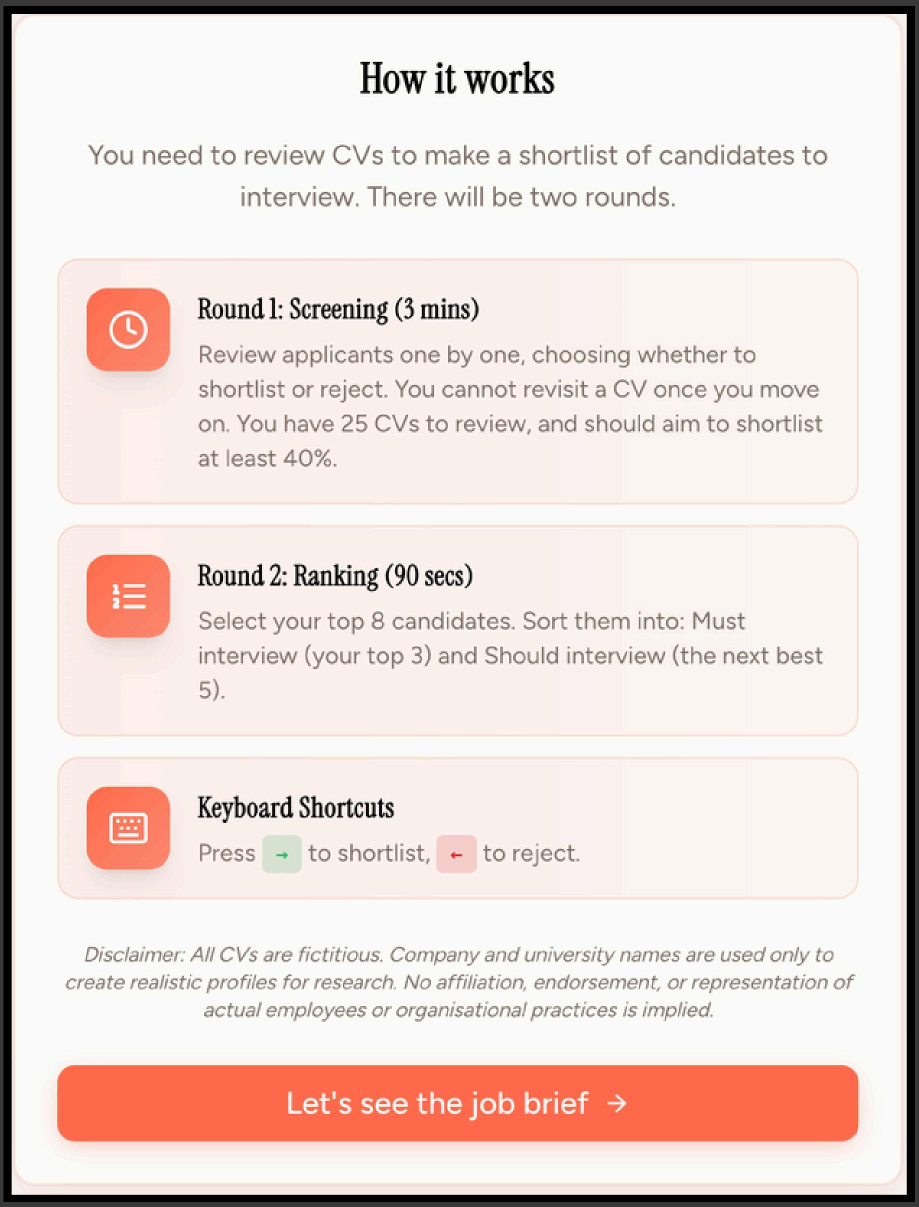


Congratulations on your fundraise!

Ready to hire your next employee?

You've received **25 applications** for your role already. You have **5 minutes** to choose which candidates to interview. Make high quality decisions as fast as you can!

You will need a laptop to complete this exercise.



How it works

You need to review CVs to make a shortlist of candidates to interview. There will be two rounds.

Round 1: Screening (3 mins)

Review applicants one by one, choosing whether to shortlist or reject. You cannot revisit a CV once you move on. You have 25 CVs to review, and should aim to shortlist at least 40%.

Round 2: Ranking (90 secs)

Select your top 8 candidates. Sort them into: Must interview (your top 3) and Should interview (the next best 5).

Keyboard Shortcuts

Press → to shortlist, ← to reject.

Disclaimer: All CVs are fictitious. Company and university names are used only to create realistic profiles for research. No affiliation, endorsement, or representation of actual employees or organisational practices is implied.

The application volume crisis.

Since 2022, applications per hire have tripled.

Recruiters report 3-6x more applications than before. Every minute on LinkedIn, 9,500 job applications are submitted.

With the increase in volume comes a fall in the ability for TAs to assess quality. AI tools now tailor CVs to job ads, generate cover letters, and answer application questions. A Yale study of 5 million cover letters found that 62% of workers had used an AI writing tool, and the correlation between cover letter quality and job offers fell off a cliff by 79%.

The old signals aren't working.

Hiring teams are rethinking their approaches. Some are raising the application bar with work samples or attempting to screen out AI-generated applications altogether. Whatever the approach, the question becomes how to do it fairly. If both CVs and work samples can be gamed with AI, then one answer is to find ways AI can help address the problem it's created.

Mark Egan, Principal Research Advisor at the Behavioural Insights team, describes his thinking,

"…We could potentially use AI to rigorously assess CVs at scale, providing another datapoint alongside the work samples and written tests we set. The key is ensuring we're measuring what matters, actual capability indicators, rather than reinforcing conventional credentialism…"

So TAs and hiring managers face an impossible set of options: allocate more resources you don't have? Screen faster and risk mistakes? Review fewer applications and miss good candidates? Or delegate to automated systems that might not be fit for purpose?

Most teams are doing a mixture of the above but uptake of AI is not slowing. A Workday report said 77% of organisations plan to increase AI use in recruitment this year. The question then becomes "which tools, and how?"

03
Regulation is here. Now.

There is no one regulation.

It depends where you operate, what you do and how you act. But there's an unmistakeable direction of travel across four big areas: more transparency, bias testing, human oversight and audit trails.

Jurisdictions are tightening the rules. The EU classifies all employment-related AI as "high-risk" under the AI Act, with full enforcement likely from August 2026. The UK takes a lighter-touch approach through ICO guidance, but the principles are the same. In the US, NYC and California already enforce AI-specific hiring rules, with Illinois and Colorado following in 2026.

The critical point is that liability falls on employers, not vendors.

In GDPR speak, you are the data controller, who will appear at the tribunal if something goes wrong. Your AI provider is the data processor.

While primary liability rests with the employers, evolving case law suggests vendors may share ‘agent’ liability. In California, a lawsuit alleges Eightfold violated the Consumer Credit Reporting Agencies Act by using candidates' data to score them without obtaining their consent. The ongoing Mobley v. Workday case may establish that AI vendors can be held directly liable as "agents" of employers. Underlying anti-discrimination law applies to employment decisions whether made by humans or algorithms. Colorado and the EU are starting to regulate both developers and deployers with distinct obligations for both.

Whoever you are, if you're using AI in hiring, you need to know what it's doing and be able to prove it's fair.

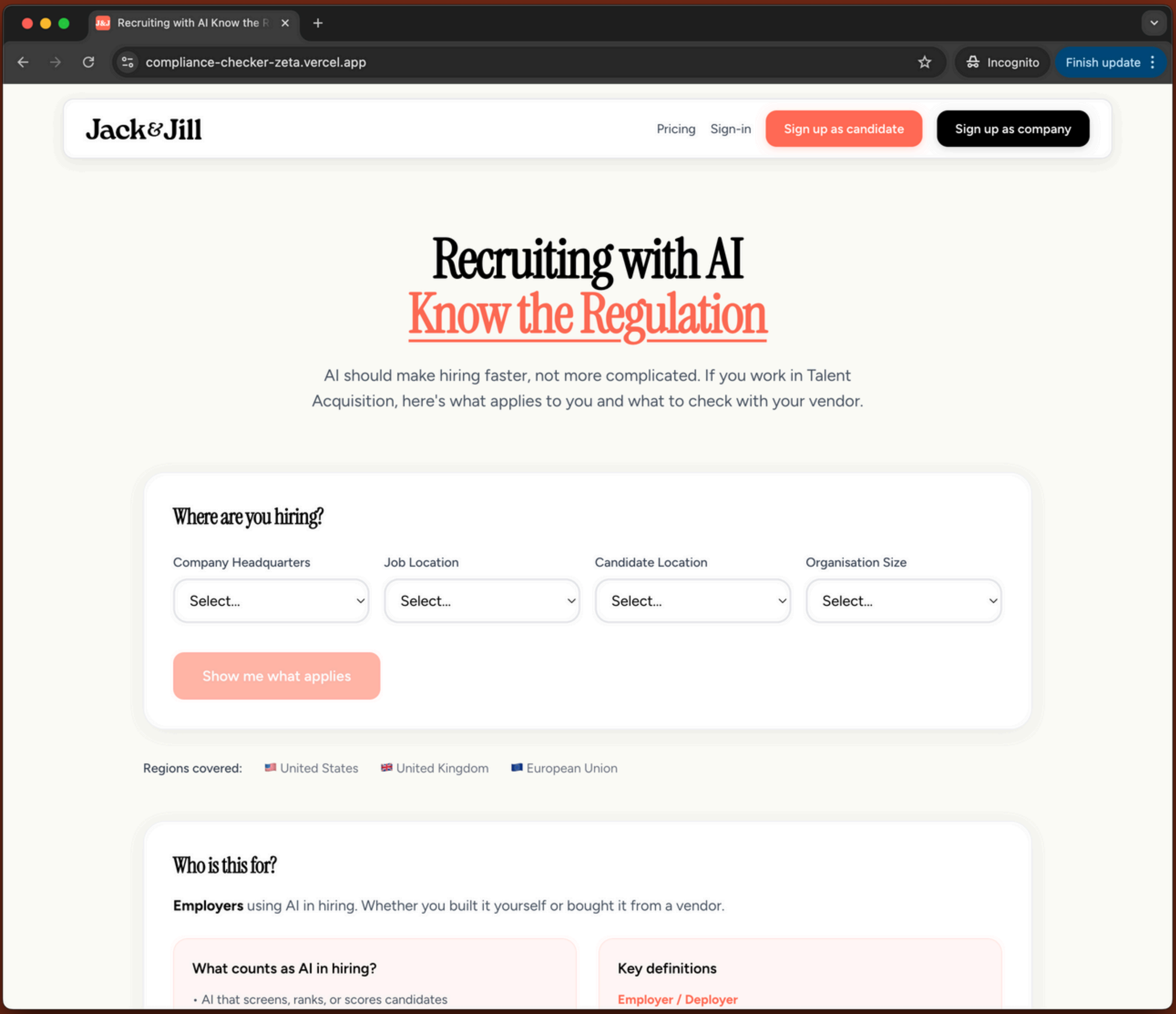
The rules.

Here's a summary of the most important regulations affecting talent leaders hiring in the UK, US and EU.

Jurisdiction	Regulator	What Employers Need to Know	Potential Fines	Status
UK 🇬🇧	ICO	▪ Complete a Data Protection Impact Assessment (DPIA) before buying software. ▪ Explain the logic/impact to candidates; they must be able to challenge decisions.	Up to £17.5M or 4% of turnover	✅ In effect
California 🇺🇸	FEHA	▪ Be ready to prove the tool is job-related and necessary if legally challenged. ▪ Keep all selection records for at least 2 years.	Damages & Injunctions	✅ In effect
EU 🇪🇺	AI Act	▪ Zero tolerance for prohibited practices (e.g., emotion recognition). ▪ You must explicitly inform candidates before AI is used.	Up to €35M or 7% global turnover	⏳ Coming: Aug 2026
NYC	LL 144	▪ Must notify candidates 10 days before use, specify what's assessed and offer alternative process. ▪ Annual third-party audit, selection rates by sex, race, ethnicity + intersections.	Up to $1,500 per violation	✅ In effect
Illinois	AI Interview Act	▪ Must notify candidates AI will analyse their interview. ▪ Explain how the AI works and what characteristics it evaluates.	Private right of action	✅ In effect
Colorado	SB 24-205	▪ Have a risk management policy for use of high-risk AI. ▪ Complete an impact assessment before deploying.	Up to $20,000 per violation	⏳ Coming: Jun 2026

Know the regulations that apply to you.

For a comprehensive list, we've built a dynamic regulation checker specially for readers so you can see exactly the ones that affect you.



Recruiting with AI
Know the Regulation

AI should make hiring faster, not more complicated. If you work in Talent Acquisition, here's what applies to you and what to check with your vendor.

Where are you hiring?

Company Headquarters	Job Location	Candidate Location	Organisation Size
Select...	Select...	Select...	Select...

Regions covered: 🇺🇸 United States 🇬🇧 United Kingdom 🇪🇺 European Union

Who is this for?

Employers using AI in hiring. Whether you built it yourself or bought it from a vendor.

What counts as AI in hiring?

- AI that screens, ranks, or scores candidates

Key definitions

Employer / Deployer

04
How to use AI to hire, and to do it fairly.

There's no silver bullet. Bias is too complex and context-dependent for that. But we can design systems that mitigate known failures, are transparent and measure how they perform.

Four principles guide that approach.

1. Remove demographic signals before evaluation

If names influence outcomes (and the research shows they do), strip them before assessment. Automated redaction removes obvious signals such as names, photos, pronouns, dates of birth. For non-text inputs, like voice, transcribe it so the system evaluates what was said, not how it sounded. Limiting what the model sees reduces the scope for bias to creep in.

But demographic information can also surface in subtler ways. Society memberships, languages spoken, extracurriculars, location, career gaps, to name a few. These proxy indicators are harder to catch. Some leakage is inevitable: a school name carries demographic signal even after the candidate's name is removed.

Eduard Schikurski, CTO of Warden AI, explains this problem and how you can stress test it.

"... Redaction can reduce direct identifiers, but it cannot fully eliminate proxy indicators of protected characteristics. An approach to bias audits includes counterfactual testing, systematically stress-testing model outcomes by modifying or removing proxy information (for example, names, gendered terms, or age signals) to determine whether those proxies materially influence decisions ..."

2. Assess against pre-set criteria, consistently

Define what you're looking for before you start screening. Evaluate every candidate against the same criteria. Research on structured interviews supports this process: predefined criteria and consistent evaluation outperform unstructured approaches.

In practice, it's hard to sustain manually. Hiring managers can fill out scorecards, but there's subjectivity in how criteria are interpreted. There’s limited accountability that scoring meaningfully contributes to decisions and under time pressure the system is sometimes dropped altogether.

Automated systems can enforce this consistently. Hiring managers define criteria upfront with clear indicators of what "good" looks like, the system evaluates every application against those criteria, and top-scoring candidates surface based on those scores.

This also means constraining where decision making is delegated to AI. A model asked to "pick the best candidates" has room to apply its own patterns. A model asked to "score this candidate's experience against criterion X" has less. Narrow, specific tasks reduce the surface area for bias to emerge.

Too narrow though and you risk encoding bias against certain groups and filtering out great performers through false negatives. The best way to avoid this is to make a rubric that doesn’t encode subjective judgements that correlate with demographics (e.g., “culture fit”) and that you monitor your system continuously (more on this later).

> “...make a rubric that doesn’t encode subjective judgements that correlate with demographics...”

3. Make reasoning transparent

Transparency is no longer optional. Regulators across the US, EU and the UK demand it.
The good news is this is a natural consequence of structured evaluation. When decisions are made against a rubric, reasoning can be captured at each step.

But not all explanations are equal. The key distinction to design for is whether reasoning drives the decision or gets generated after the fact. A system that reasons as it decides evaluates each candidate against defined criteria, scores them, and surfaces top performers based on those scores. A system that explains after deciding first selects candidates, then generates rationale. The second produces explanations, but they're retrofitted: the reasoning didn't shape the outcome.

Monitoring catches what design choices miss.

True transparency uses reasoning as the decision mechanism, not post-hoc justification. This creates audit trails for regulators and enables genuine, specific feedback for candidates based on how they actually scored. This is a key feature to demand of your AI provider.

4. Monitor continuously to catch what design misses

Even with Principles 1-3, bias can persist. Proxy signals leak through redaction and criteria themselves can encode bias. Continuous monitoring – monthly ideally, or quarterly for smaller companies – catches what design choices miss, but only if paired with meaningful human oversight and clear remediation processes.

First, you need to check for failure.

Where demographic information is available (e.g., opted-in for analysis purposes only), patterns become visible. If particular criteria consistently disadvantage specific groups, that shows up in the data. Similarly, if model scores drift over time, candidates report feedback that doesn't match their qualifications, or human reviewers frequently override automated recommendations, you probably have a problem.

You then need to be accountable and act on these findings. Revising criteria, adjusting weightings, reviewing affected candidate cohorts, or pausing screening while you investigate are just some of the ways to respond to what you find.

Final hiring decisions must involve human judgment; it's good practice and a regulatory requirement.

05 Beyond screening: finding and supporting candidates fairly.

Beyond these four principles focused on screening candidates, we need to design for fairness for sourcing and interviewing too.

Reaching candidates who wouldn't otherwise apply

Demographic factors shape who applies. Women apply to 20% fewer jobs than men despite similar browsing behaviour and tend to apply only when meeting a higher proportion of stated requirements. Those with caring responsibilities have less time for job-seeking so are disadvantaged.

Direct outreach (finding and inviting qualified candidates rather than waiting for applications) widens the pool to mitigate for this. At scale, automation makes this feasible in a way manual headhunting cannot, such as larger networks, consistent criteria, less reliance on existing relationships.

Helping candidates put their best foot forward

Candidates can only articulate what they recognise as valuable.

Strong candidates, especially those from marginalised backgrounds, don’t lack skills. They lack what Nina Slingsby, CEO of OAHA, calls the “translation layer”, which tells them what employers really mean by the skills listed in a job ad, and what good evidence looks like. They need clear explanations,

“...By explaining the skills we’re hiring for in a straightforward, inclusive manner and giving examples of what those skills can look like in real life (work, caring responsibilities, community roles, volunteering, part-time jobs), we can help people map their experience to the role without lowering the bar..."

This isn't about confidence. It's an information and framing gap. AI can help provide this translation layer with clear skill definitions, examples of evidence across different contexts, and prompts that help candidates surface relevant experience they may not realise counts.

This makes skills-based hiring easier to understand and fairer to access.

06
Conclusion.

The decision to use AI in hiring has already been made. The question is which tools to choose and how to deploy them.

General-purpose models aren't built for this. Indeed, many readily available models will straight-up refuse to assess CV data because of known limitations.

They carry known biases, lack appropriate guardrails, and offer little in the way of transparency or audit trails. Using them to screen candidates exposes you to regulatory risk and puts your candidates at risk of unfair treatment.

But recruitment-specific AI, designed with bias mitigation and transparency in mind, offers a different path.

The approach in this paper to strip out demographic signals, abide by pre-set criteria, make reasoning transparent and evaluate at scale are starting points for measurable improvement over time.

None of this guarantees perfect outcomes. Anyone claiming that should be carefully scrutinised. But it does create infrastructure that's more consistent, more transparent, and easier to improve than purely human processes at scale.

07
Your checklist.

Based on the principles in this paper, here's your checklist as a TA leader.

Actions you need to take:

- Define clear hiring criteria upfront before screening starts
- Know your regulatory requirements - use our tool
- Maintain audit trails for all AI-assisted hiring decisions
- Keep human oversight on final hiring decisions

Questions to ask vendors:

- What demographic signals does your system see, and how do you test whether they influence outcomes?
- How is reasoning captured — does it drive decisions, or explain them afterwards?
- How often do you carry out audits, and who conducts it?
- Can you show me your most recent bias audit results?
- Can you provide specific, meaningful feedback to candidates?

If they can’t give satisfactory answers to these questions then it’s time to look elsewhere.

The tools exist to do this. You need to work out whether you're using them.

If you want to discuss anything in this paper or help make our system better please email us at legal@jackandjill.ai

References.

ACLU/Public Justice. "Complaint Against HireVue." 2025.

Ashby. "Candidate Application Volumes Triple Post-ChatGPT." 2024.

Behavioural Insights Team. "Gender and Job Applications." 2022.

Bertrand, M. & Mullainathan, S. "Are Emily and Greg More Employable than Lakisha and Jamal?" 2004.

Brookings Institution. "Childcare and Labour Market Participation." 2022.

Chen et al. "Systematic Bias in LLM Resume Screening." 2025.

CIPD. "Resourcing and Talent Planning Survey." 2015.

Cole et al. "Inter-rater Reliability in Resume Screening." 2009.

Financial Times. "AI-Generated Applications Flood Job Market." 2024.

Kline, P., Rose, E. & Walters, C. "Systemic Discrimination Among Large U.S. Employers." 2022.

LinkedIn. "Gender Insights Report." 2019.

LinkedIn. "Platform Statistics." 2025.

Lippens et al. "Disability, Age, and Physical Appearance Discrimination in Hiring." 2022.

Mobley v. Workday. Ongoing litigation. 2025.

Park, J. & Oh, I. "Meta-analysis of Gender Discrimination in Hiring." 2025.

Pedulla, D. "Race and Unemployment Stigma Interactions." 2018.

Quillian et al. "Meta-analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time." 2017.

Reuters. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." 2018.

Rivera, L. & Tilcsik, A. "Gender and Class Signal Interactions." 2016.

Sackett et al. "Structured Interviews and Hiring Quality." 2022.

Starcke, K. & Brand, M. "Stress and Cognitive Control in Decision Making." 2016.

Webster, K. "Contextual and Intersectional Bias in AI Hiring Tools." 2025.

Wilson, J. & Caliskan, A. "Bias in Large Language Models for Hiring." 2024.

Workday. "State of AI in HR Report." 2024.

Yale/Freelancer. "Impact of AI Writing Tools on Cover Letter Quality and Job Outcomes." 2024.