

Developing and assessing a scale for measuring human–AI alignment in hiring decisions

Andres Abril

Inter-American Development Bank

Marcos A. Rangel

marcos.rangel@duke.edu

Duke University

Wladimir Zaroni

Inter-American Development Bank

Article

Keywords: human–AI alignment, algorithmic hiring, item response theory, Turing test, labour markets, large language models

Posted Date: October 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10630904/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Developing and assessing a scale for measuring human–AI alignment in hiring decisions

[Author names removed for anonymous review]

[Affiliations removed for anonymous review]

Abstract

We propose a *Labour Market Turing Test* (LMTT) to evaluate the alignment between human and artificial intelligence (AI) decision-making in hiring. Using a randomised lab-in-the-field experiment, we compare referral decisions made by professional recruiters with those produced by large language models (LLMs). We introduce *machina*, a latent trait that places human and AI entities on a shared scale according to how closely their referral descriptors mimic a reference-AI recruiting decisions. The distribution of this trait differs markedly between humans and AI teams in their means and variances. When AI systems are tuned to impersonate human-like characteristics, these differences narrow but do not vanish, revealing persistent unexplained heterogeneity. Indeed, hybrid models that more heavily weight random rather than “rational” AI choices more accurately reproduce observed human decision patterns. We further show that the degree of alignment between humans and AI varies systematically with job complexity and the subtlety of candidate differences. Together, these findings suggest that perfect imitation of human judgment requires capturing not only average tendencies but also the diversity of human reasoning. We argue that LMTT can provide a generalisable framework to quantify and interpret context-specific alignment between human and AI decisions in labour markets, offering a basis for evaluating complementarities and trust in algorithmic hiring systems.

Keywords: human–AI alignment; algorithmic hiring; item response theory; Turing test; labour markets; large language models

1 Introduction

Screening and selection processes influence nearly every aspect of modern life, from personal relationships to key institutional decisions in healthcare, criminal justice, education and labour markets. Once driven exclusively by human judgment, these decisions are now increasingly guided, or even entirely undertaken, by artificial intelligence (AI). In the present article we devise a framework to measure and document the contrast between those two types of decision-making entities across multiple labour market contexts.

We argue that the screening processes within labour markets provide a particularly compelling case study for understanding the divergence between human and algorithmic decision-making. With some job postings attracting hundreds or even thousands of applications, companies increasingly rely on AI to screen applicants before those reach human reviewers. In fact, according to recent press coverage (Kaufman, 2025), 48% of hiring managers now use some form of AI-

driven screening, and the human resources industry forecasts a compound annual growth rate of 6.1% in AI utilisation through 2030. This rapid diffusion is consistent with broader evidence that AI technologies are reshaping employment levels and the structure of labour demand across economies (Shen and Zhang, 2024). In this high-stakes context, the rapid integration of AI into hiring represents a profound yet sometimes inadvertent transformation in recruitment practices, requiring a critical examination of how these new systems may choose differently from their human counterparts. Moreover, given the significant impact hiring decisions have on people’s lives, AI is poised to replace humans as major gatekeepers in mediating access to economic opportunities. Since the portrayal of AI as an objective decision-maker often conflates the systematic nature of an algorithm with true impartiality, we call attention to this phenomenon as having far-reaching implications not only for efficiency but also for equity.

We make the case that the understanding of the consequences of AI utilisation in labour market screening cannot advance without a methodology to quantify (metric) not only **if** AI decisions look different than the ones made by an average human, but also from a careful assessment of: (i) how distinct these decisions are from each other; (ii) how heterogeneous these distances are within the human population; (iii) what are the main drivers of such heterogeneity; and (iv) in which contexts these differences are most salient. In other words, before evaluating the consequences of delegating screening decisions to machines, there is a need to devise not only solid strategies to assess the existence but also to quantify the differences in AI and human decision-making, which capture both individual heterogeneity and the influence of contextual elements (complexity of labour markets).

While prior research has compared algorithmic versus human performance at screening (Will et al., 2023; Hoffman et al., 2018; Li et al., 2020) and reviewed the sources of algorithmic bias in AI-enabled recruitment (Chen, 2023), few studies have implemented controlled experiments to measure alignment between human and AI judgments directly to identify the conditions under which such alignment is more or less likely to break down. Using primarily collected data from a lab-in-the-field randomised control trial with real recruiters in Ecuador, we evaluate how AI recruiter teams perform the same candidate selection tasks relative to their human counterparts. Specifically, we measure differences in decisions when recruiters (human and AI) assess pairs of job applicants across a myriad of contexts defined by vacancy requirements, occupation, sector, and candidate profiles. Different (untrained) AI teams were constructed to vary in the degree to which they are asked to impersonate human-like traits, as defined by the profiles captured by survey instruments filled by human recruiter counterparts.

We refer to our method to systematically compare human and AI decisions as a *Labour Market Turing Test* (LMTT) framework. Leveraging psychometric techniques, we define a latent trait, *machina*, which captures how AI-like an agent’s choices (human or otherwise) appear. The method generates a unidimensional metric (or scale) of behavioural distance between each AI agent and each human recruiter in our experiment. Ultimately, it allows us to compare the decisions and behavioural patterns underlying them relative to a benchmark, offering insight into both how different the choices are and what can or cannot account for those differences along a common scale. In addition, the scale construction reveals how particular job specifications and

candidate profiles help better distinguish humans from machines.

2 Human recruiters

To evaluate hiring referral decisions, we draw on decisions undertaken by 277 freelance human resources recruiters we hired for a lab-in-the-field experiment conducted in Ecuador in 2022. The task was presented to them as a genuine job-hiring exercise, ensuring their engagement and commitment. The contact, follow-up, hiring, and payment of recruiters occurred in the second half of the year via the operation of a reputable non-governmental organisation with expertise in job search and training programmes in Ecuador.¹ The process for reaching human resources personnel involved two sampling methodologies: (i) advertisement of a job opportunity on LinkedIn and (ii) respondent-driven sampling (RDS) referrals or “snowball” sampling (Shaghaghi et al., 2011). LinkedIn respondents were motivated to participate after seeing targeted advertisements, while the RDS approach began with a small group of initial participants (“seeds” from LinkedIn) who were encouraged to refer others with similar profiles. Materials used on these approaches can be seen in Supplementary Fig. S1. Using different recruitment strategies, the study aimed to capture a diverse range of perspectives and ensure the representation of various segments of the labour market. Recruiters were remunerated competitively (according to the rates of the local market) for their time and expertise.²

Before conducting their tasks, hired recruiters also completed a detailed background survey, including demographic characteristics, educational and professional experience (with a test of their basic knowledge of the Ecuadorian labour market). The questionnaire also captured information aimed at creating a detailed profile by capturing cognitive ability (Wonderlic test), personality traits (Big Five Inventory), and self-esteem (Rosenberg scale). The Wonderlic test is a timed 12-minute assessment widely used to evaluate problem-solving skills and general intelligence (Wonderlic, 1999). The Big Five Inventory (BFI) assesses personality based on five key traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism (John et al., 1991), while the Rosenberg Self-Esteem Scale is a commonly used 10-item tool that captures individuals’ overall perception of self-worth (Rosenberg, 1965). We reproduce descriptive statistics in Table 1.

We find that human recruiters in our experiment (36% of them hired via LinkedIn contact) were 33.8 years old on average, and 29% of them were male. They were also highly educated: 93% had completed tertiary education (78% in an HR-related major/specialisation), and 1 in 5 had completed some graduate-level programme. Participants had, on average, 3 years of work experience in HR-related roles. Approximately 61% were employed at the time of participation, and 13% had some work experience outside Ecuador. Responses to labour market knowledge questions regarding minimum wage and basic labour rights indicated strong awareness, with 86% scoring in the top category. In terms of socio-emotional traits, we observe that our average

¹Name of partner organisation withheld for anonymous review.

²After manifesting interest, these recruiters were approached via email message (Supplementary Fig. S2). In this communication, recruiters were informed of the two-step process involved in the execution of their freelance task.

Table 1: Human recruiters: descriptive statistics

	Mean	Standard Deviation	Skewness	Percentile 25	Percentile 50	Percentile 75
	[1]	[2]	[3]	[4]	[5]	[6]
<i>Recruitment method</i>						
Recruited via LinkedIn (=1)	0.36	0.48	0.56	0.00	0.00	1.00
<i>Demographics and SES</i>						
Male (=1)	0.29	0.46	0.91	0.00	0.00	1.00
Age (years)	33.80	7.20	1.47	29.00	31.00	37.00
Medium-high SES (=1)	0.10	0.31	2.58	0.00	0.00	0.00
High-SES (=1)	0.07	0.26	3.31	0.00	0.00	0.00
<i>Education</i>						
Postgraduate degree (=1)	0.20	0.40	1.51	0.00	0.00	0.00
College degree (=1)	0.73	0.45	-1.03	0.00	1.00	1.00
HR-related major/degree (=1)	0.78	0.41	-1.38	1.00	1.00	1.00
<i>Work experience</i>						
Work experience in HR (years)	3.07	2.27	0.92	1.00	2.00	5.00
Currently employed (=1)	0.61	0.49	-0.47	0.00	1.00	1.00
HR-related current/last job (=1)	0.91	0.29	-2.79	1.00	1.00	1.00
HR experience abroad (=1)	0.13	0.34	2.15	0.00	0.00	0.00
High knowledge of Ecuadorian market (=1)	0.86	0.35	-2.07	1.00	1.00	1.00
<i>Socio-emotional and Cognitive Scores</i>						
Neuroticism (0-48)	30.71	3.55	-0.11	28.00	31.00	33.00
Extroversion (0-48)	32.34	3.75	-0.12	30.00	32.00	35.00
Openness (0-48)	29.04	3.74	-0.12	26.00	29.00	32.00
Agreeableness (0-48)	31.76	3.89	0.01	29.00	32.00	34.00
Conscientiousness (0-48)	30.18	3.79	0.22	28.00	30.00	33.00
Rosenberg Self-esteem (10-40)	34.19	5.79	-1.52	32.00	36.00	38.00
Wonderlic Cognitive test (1-46)	16.75	4.49	-0.01	13.00	17.00	20.00

Note: Statistics based on 277 unique human recruiter observations.

recruiter scored high on the Rosenberg instrument, which is considered to be within the normal range (above 30) in the literature. The same was true for most of the Big Five constructs, with the most obvious counterexample being Openness (which sat at mid-range levels). Finally, cognitive scores revealed low performance, suggesting that our average recruiter may handle routine work only with supervision and, rather surprisingly, may not be well-suited to jobs involving advanced reasoning and learning.

When we compare these data against national-level labour force data from Ecuador’s household survey (ENEMDU, 2021–2023), we see that while not representative, the sample displays meaningful demographic skews: participants were younger on average (33.8 years vs. 37.7 years nationally), more likely to be female (71% vs. 58.3%), and more likely to hold a college degree (93% vs. 80.3%). These differences likely stem from recruitment channels, particularly LinkedIn, which tends to attract younger and more formally educated professionals in the Ecuadorian labour market (Zanoni et al., 2024a).

3 Job recruiting tasks

Using a web-based platform mimicking standard human resources software (see Supplementary Figs. S3 and S4), the human recruiters were tasked with referring best-fit candidates for each job vacancy with which they were presented, each one with exactly two applicants. To ensure natural engagement without priming, the task was framed as a professional consulting assignment

aimed at testing the platform’s reliability. While all recruiters were assigned the same set of 10 job vacancies, each had candidate pairings randomly assigned. As further detailed below, each pairing had candidates roughly matched on productivity-related traits (e.g. education, experience, age) that differed in identity markers (themes). The latter should, at least in principle, be irrelevant to candidate selection because they do not explicitly represent productive attributes among the jobs we were hiring for.³

3.1 Vacancies

The vacancies in the experiment spanned various sectors with differing required/desired education, experience, specialisation and responsibilities: (1) Sales Representative, (2) Warehouse Assistant, (3) General Janitorial Services Assistant, (4) Maintenance Technician, (5) Technical Project Manager, (6) Production Supervisor (Manufacturing), (7) Software Developer, (8) Systems Engineer, (9) Certified Public Accountant (CPA), and (10) Call Centre Operator. Details of each vacancy are presented in Supplementary Table S1.

3.2 Identity marker themes

Evaluations of employability were based on productivity markers provided on the web platform. In some cases, the productive attributes were aligned with the identity marker. For instance, the institution where education was acquired matched one’s country of origin. From the randomisation of identity markers, in addition to a placebo exercise, a total of three thematic categories were used: gender, sexual orientation, and immigration status. In some cases, these were represented through multiple variations. The final design included 10 unique paired-CV versions for each of the vacancies listed above: (1) non-immigrant straight woman vs. non-immigrant straight woman; (2) non-immigrant straight woman vs. non-immigrant straight man (3 versions); (3) non-immigrant gay man vs. non-immigrant straight man (2 versions); (4) non-immigrant lesbian vs. non-immigrant straight woman; (5) immigrant straight man vs. non-immigrant straight man; and (6) immigrant straight woman vs. non-immigrant straight woman (2 versions).

3.3 Summary

The final experimental structure is described diagrammatically in Supplementary Fig. S5. We henceforth refer to each of these 100 vacancy-candidate pair variants as a “trial”. These trials are described in Supplementary Table S2.

Ultimately, each recruiter was tasked with a unique random sequence of 10 trials (one of each vacancy type and one of each identity theme variant, necessarily), with the random allocation resulting in each specific pre-designed trial being seen by 17 to 40 unique human recruiters.

³This experiment has been previously used to examine patterns of unequal treatment in hiring decisions related to an applicant’s attributes (Fabregas and Zanoni, 2025; Zanoni et al., 2024a,b).

4 AI recruiters

We developed four distinct AI-recruiter teams using ChatGPT (version GPT 4o-mini), each comprising 277 agents, mirroring the structure and sample size of the human recruiter sample. In developing the AI recruiters, we adapted the method of cognitive interviewing (Willis, 2005), which is commonly used in survey research to assess how respondents understand questions, to evaluate whether AI agents correctly interpret our prompts. Our method aligns with approaches in prompt engineering, such as the “Cognitive Verifier Pattern”, which guides AI behaviour through structured interactions to ensure comprehension and clarify ambiguity. This technique is consistent with Human-AI Interaction (HAIC) principles that emphasise trust, interpretability, and adaptability in AI systems (Liu et al., 2019; Vanderbilt University Generative AI Initiative, 2025; Turner, 2025; Microsoft, 2025). Through structured dialogue, we assessed the AI’s understanding of the task, its interpretation of human-like traits some of them are asked to impersonate, and the semantic structure of its outputs. This application of cognitive interviewing in prompt engineering offers a systematic framework for testing and refining AI behaviour in high-stakes decision contexts such as hiring. Moreover, it emulates a conceptual bridge between human cognition and algorithmic reasoning, one of the core inquiries of our research agenda.

Figure 1 presents a summary infographic explaining the creation of the AI recruiter teams and their contrast with human recruiters.

Recruiters	Hire/Impersonate the Recruiter	Select the Candidate
Robot 		- Read and interpret the HTML program of the web page to evaluate candidates - Select Jobs sequentially from a list of vacancies - For each selected job, read and interpret the information about the candidates. - Make a candidate referral
Avatar 	As humans did, AI were asked to <i>read and interpret</i> - The LinkedIn post or RDS email sent - The job instructions sent by email - The HTML program of the registration page - The recruiters’ characteristics Then, they were asked to <i>impersonate</i> the recruiters’ characteristics	- Read and interpret the HTML program of the web page to evaluate candidates - Select Jobs sequentially from a list of vacancies - For each selected job, read and interpret the information about the candidates. - Make a candidate referral
Clone 	As humans did, AI were asked to <i>read and interpret</i> - The LinkedIn post or RDS email sent - The job instructions sent by email - The HTML program of the registration page - The recruiters’ characteristics Then, they were asked to <i>impersonate</i> the recruiters’ characteristics	- Read and interpret the HTML program of the web page to evaluate candidates - Select jobs in the same order as human recruiters did. - For each selected job, read and interpret only the candidate information viewed by human recruiters. - Make a candidate referral
Human 	- Contacted via LinkedIn or RDS. - Read and accepted the job offer. - Accessed the registration page. - Completed a questionnaire with personal and profile information.	- Read Instructions of the selection page. - Select Jobs sequentially from a list of vacancies. - For each selected job, read and interpret the information about the candidates. - Make a candidate referral.
Cognitive Interviewing (Prompt Examples)	- Analyze all of the recruiter’s characteristics and identify how they reflect their experience, skills, personality, and technical knowledge. - Summarize how these characteristics might influence their ability to select candidates. - Describe the recruiter’s selection style based on their full profile (e.g., focused on technical experience, interpersonal skills, or a balance of both). - Based on the recruiter’s characteristics you previously discussed and the aspects of the email you identified at the beginning of the process, you will now assume their role.	- Describe what type of information you believe you need to analyze in order to make decisions at this stage. - Explain your main task when analyzing vacancies and candidates, outline the specific steps you will follow, and mention any additional criteria or information that will help you make decisions. - Summarize key traits of each candidate, compare strengths and weaknesses, identify decisive factors, select the best fit, and briefly justify your choice based on the recruiter, the vacancy, and the candidates

Figure 1: Human and AI recruiter teams

All our AI agents followed the standardised protocol of selecting one candidate per trial, resetting memory after each decision, and evaluating the same 10 vacancies and candidate profiles as their human counterparts without feedback. We did, however, impose that no memory was to be carried across trials. Importantly, we did not train the AI recruiters on any human

decision-making data. Instead, aside from the prompts designed to guide their behaviour, we relied on their inherent, pre-existing algorithmic knowledge. Our four AI teams differed only in the degree of instruction they received before executing the 10 trials within their screening task. Each team was provided access to different types of instructional information, organised into four blocks: (a) instructions on the task to be performed; (b) required “impersonation” of background characteristics of human recruiter counterparts; (c) an imposed sequence of decisions during platform use; and (d) a limitation on the set of environments in the web platform displayed during decision process (defined to correspond to each human counterpart’s sequence and, therefore, observed candidate-information set). To aid communication, we adopted the following labelling: *Robots* were only subject to (a); *Avatars* to (a), (b), and (c); and *Clones* to (a), (b), (c), and (d). Our final group, *Randomistas*, created a reference choice procedure that emulated a particular realisation of a coin toss every time a choice was required. This gradation from *Robots* to *Randomistas* illustrates a range of possible alignments with human context and behaviour.

Robots. A total of 277 Robots joined an AI team of recruiters that received minimal instructions. Robots had autonomy in choosing the order in which the 10 vacancies and corresponding applicants would be evaluated. Each of the robots assigned 10 trials could begin with technology-oriented jobs, less specialised roles, or follow a random order. Yet, because no memory was to be carried across trials, the order in which the trials are undertaken is immaterial to our analysis.

Avatars. A total of 277 Avatars joined another team of AI recruiters. Each one of the Avatars was explicitly instructed to impersonate a counterpart human recruiter in the experimental data. Impersonation is suggested by simply communicating to the algorithm the measures contained in a vector of background characteristics. This included demographics, cognitive ability scores (Wonderlic), personality trait scores (Big Five Inventory), self-esteem scores (Rosenberg scale), prior professional experience, and knowledge of the local labour market. The impersonation also included the method of recruitment (RDS or LinkedIn), understanding of the original instructions included in email correspondence from the hiring process, and the HTML content of the web platform used by human recruiters to conduct the evaluations.

While each Avatar was paired with a specific human recruiter and asked to contextualise its decisions based on detailed information about that individual, they were not instructed either on how those characteristics should influence their choices or on how human recruiters reviewed the job applications in the web platform. Avatars, just like all AI recruiters we created, were untrained algorithms.

Clones. An additional team with 277 Clones was created, each designed not only to impersonate a specific human recruiter (as an Avatar) but also to follow the same observed interaction with the platform during the evaluation task that its human counterpart followed. Unlike other synthetic recruiters, Clones were constrained to (1) view only the information that their corresponding human recruiter had elected to access when evaluating a given candidate pair. This restriction reflects the platform’s tab-based interface, where human recruiters could choose

which tabs to open for each applicant. As a result, Clones were shown only the potentially reduced subset of information actually viewed by their human counterpart. Furthermore, Clones were instructed to mirror the exact sequence chosen by their counterpart human recruiter when evaluating the job vacancies in their queue.

Randomistas. Our final team of 277 AI recruiters established a purely (realised) stochastic baseline. Each Randomista selected a candidate at random (coin-toss) in every trial to which they were exposed.

5 A Labour Market Turing Test

In order to systematically contrast human and algorithmic judgments under these conditions, we propose a scale development inspired by the classic Turing Test (Turing, 1950), and adapted to personnel hiring: a *Labour Market Turing Test* (LMTT). The LMTT formalises the comparison within and between human and AI recruiters by asking whether and how AI and human sequences of referral choices can be distinguished from one another at the individual level. Despite the distinct set of vacancy-CV contrasts these individual recruiters are subject to, this method still allows us to assign a scale/metric to the differences that may emerge. This test’s conceptualisation is based on the premise that while both humans and machines must navigate similar constraints, they likely do so in distinctive manners. For example, humans may draw on tacit knowledge and socio-emotional reasoning, while AI entities may leverage structured data, information management capacity, and algorithmic rules.

We implemented the LMTT using binary choice tasks between synthetic candidates, building on the tradition of discrete choice experiments in economics (Eriksson et al., 2017) and paired comparison protocols from machine learning (Amershi et al., 2019). This binary format mirrors real-world short-listing decisions, reduces complexity for human evaluators, and facilitates controlled comparison across recruiter types. As described above, for each of 10 job openings, both human and AI recruiters were presented with a particular sequence of candidate pairs and asked to select exactly one as the most qualified for the job (see diagram in Supplementary Fig. S6). We highlight, however, the potential heterogeneity across choice pairs. In fact, we detect substantial variation in the ambiguity of choices within each trial. Supplementary Figure S7 portrays the “margin-of-victory” for the preferred candidate in each of the 100 trials, considering all team-specific recruiters seeing the same candidate-pair/vacancy (we display humans and robots). We find, for example, that even robots with the same set of instructions can choose different candidates when exposed to the same decision. We also report that the agreement level within human teams in general is, albeit not surprisingly, smaller.

The binary structure of the screening tasks aids the contrast across multiple recruiters and multiple ambiguity contexts we set out to develop here. To do so, we propose a model-based framework that places all individual decision-making entities, human or AI, on a common scale. Specifically, we draw from a long tradition in psychometrics and use a two-parameter Item Response Theory (IRT) model (Birnbaum, 1968; Lord and Novick, 1968). IRT, also known as

latent trait analysis, corresponds to a set of statistical models that describe the relationship between an agent’s response to exam/questionnaire items and an unobserved latent trait, such as proficiency in mathematics or prevalence of a socio-emotional trait. These methods have also been used in the measurement of wealth/poverty, support for economic inequality, and physiological dysregulation (Liu et al., 2020; Wiwad et al., 2019; May, 2006), for example. Here, we employ this framework by treating a screening trial as an item in an exam, which allows us to infer each recruiter’s underlying latent trait based on their pattern of responses to job-candidate comparisons, even when the sequence of candidates being evaluated by a given recruiter is different from the one evaluated by the next (all that is needed is a partial overlap in choices).

More specifically, our IRT model assumes that there exists an unobservable latent (and conceptually sound) trait, which we call *machina*, that cannot be directly measured, unlike observable characteristics such as human age or ChatGPT’s token limit. However, within our framework, *machina* can be indirectly assessed by comparing each recruiter’s referral choices to a reference decision that represents the prototypical behaviour of a representative AI agent. The choices of the representative AI agent serve as an exam answer key, providing a benchmark for what a perfectly machine-like response would entail—much like a mathematics exam’s answer key represents the single correct solution. In what follows, we refer to this benchmarking recruiter as *Skynet*, conceived to be representative of pure machine-like consensus behaviour. Programmed without contextual framing, *Skynet* is defined by the majority-choice candidate from each pair of applicants reviewed by the AI entities we refer to as *Robots*. That is, we define the selected candidate in each trial as the one *Robots* chose more than 50% of the time after evaluating the applications. These choices play the role of reference for comparison in what we examine below.

Skynet, therefore, was neither explicitly instructed to ignore identity-related information nor optimised for fairness. Rather, it reflects how a general-purpose AI system behaves when making referral decisions based solely on prompts that prioritise candidate suitability without any additional targeted constraints or training. Although identity markers were not emphasised in its instructions, *Skynet* may still attend to them implicitly if its underlying training data encoded patterns linking identity to job performance or selection. In fact, it is crucial to distinguish between *Skynet*, which serves as a benchmark for prototypical machine-like behaviour reflecting inherent algorithmic priorities, and the concept of fairness in these high-stakes decisions. This is because while *Skynet* provides a measure closer to algorithmic consistency, it does not, by itself, represent a normative ideal or an inherently fair outcome.

Even though our contrast between each human recruiter and *Skynet* can be conducted by simply computing the share of human choices that conform with *Skynet*’s across 10 trials, we insist that this may lead to an incomplete analysis because no two human recruiters see the same set of trials. Before advancing our model, however, we report in Supplementary Figure S8 the variability in agreement rates across the 277 human recruiters in our experiment. We find, for example, that most humans agree with *Skynet*’s choices somewhere between 50 and 70% of the time.

In order to address the heterogeneity across trials and across raters/recruiters we employ a model that, without loss of generality, characterises individual choices by denoting $(c_1 > c_2)_j^S$

as *Skynet*’s choice of candidate c_1 over c_2 in trial j . Then, for recruiter i , the probability of matching *Skynet*’s preferred pick is explicitly modelled as:

$$\Pr[(c_1 > c_2)_j^i = (c_1 > c_2)_j^S] = \frac{1}{1 + \exp\{-\alpha_j(\mu_i - \beta_j)\}}, \quad (1)$$

where μ_i represents the recruiter’s latent trait of interest. The higher the levels of this trait, the more likely recruiter i is to select the same candidate as the one chosen by *Skynet* in any given trial. In discussing our findings below, we rescale the latent scores so that *machina* is expressed in the relative units corresponding to the human distribution of the trait (z-score). The rescaled score, therefore, indicates how far a recruiter lies above or below the human average in standard-deviation units. A value of zero corresponds to the average human recruiter, positive values indicate recruiters who are more machine-like, while negative values correspond to recruiters who are less machine-like than the average human. This provides a full characterisation of heterogeneity across recruiters in their intrinsic propensity to agree with the reference machine.

As we pointed out before, this framework also allows for the identification of trials in which the potential agreement between individual recruiters and *Skynet* is more or less common. This heterogeneity across trials is captured by two meaningful parameters. The first parameter we refer to as the *location* parameter (β_j). This is analogous to the “difficulty” parameter in standard IRT models, which measures at what level of proficiency test takers start having more than a 50/50 chance to get the right answer in a given question. Here, however, our location parameter places a given trial on the latent trait continuum, indicating the trait level at which recruiters are equally likely to agree or disagree with *Skynet*’s choice. We convert this parameter using the same strategy used above for the latent trait, so that high values of this parameter indicate that only individuals with very high levels of *machina* would ultimately choose like *Skynet* when confronted with such a trial. To aid communication, however, we report the *location* parameter of a trial as a percentile in the human distribution of *machina*. In other words, our location parameter indicates the proportion of the human population for which agreement with *Skynet* becomes more likely than not. Therefore, our modelling exercise reveals in this way the heterogeneity in screening trials’ (contextual) agreement levels between humans and the reference machine.

The second parameter captured by our technique is conceptually equivalent to the “discrimination” parameter found in standard IRT analyses. However, to avoid conceptual confusion, particularly given the original field experiment study’s focus on unequal treatment of disadvantaged populations in hiring, we refer to this as the *differentiation* parameter (α_j). This parameter ultimately reflects how effectively a given item distinguishes between individuals at different points on the *machina* continuum. In practice, it measures how much the agreement level between a given recruiter and *Skynet* changes as we move along the *machina* trait spectrum. Relatively higher differentiation values indicate that a trial is very sensitive to differences in the *machina* trait across recruiters, and therefore better at differentiating the machine-like from the non-machine-like behaviour. In other words, this parameter defines how sharp the boundary is between those humans who likely and unlikely select the *Skynet* candidate in a given trial.

Intuitively, our trials are the contexts through which the *machina* trait (behaviour similarity with *Skynet*) is expressed or revealed in observable behaviour (the job candidate selected). Therefore, every trial can be seen as a kind of mechanism of revelation: it translates the latent trait into something observable. A low-differentiation item is a context in which similarity to the performance standard (*Skynet*) fails to materialise clearly in observable behaviour. The person’s true standing relative to the standard exists (*machina*), but the behavioural outcome in that trial is simply too weak to diagnose it. If we borrow from the economics intuition, the revelation principle says that mechanisms should be designed so that participants’ private information (here, latent trait) is truthfully revealed through their actions. High-differentiation trials are, therefore, “truth-revealing mechanisms”, while low-differentiation trials are “inefficient mechanisms”, where the mapping from latent similarity to observed choice of candidates is weak or distorted.

In sum, our LMTT is carefully designed to highlight the elements we deem important in measuring and ultimately understanding the trade-offs between the use of AI in hiring decisions. Its estimation procedure involves an expectation-maximisation (EM) iterative algorithm solution (Newton–Raphson). Its implementation yields a score which represents the recruiter’s latent trait of interest (*machina*). The higher the levels of this trait, the more likely a generic recruiter i is to select the same candidate as the one chosen by *Skynet* in any given trial j . At the same time, every trial (or context) j can be classified along its *location* and *differentiation* dimensions.

6 Results

6.1 Trial-level contextual characterisation

As discussed above, our IRT model also yields information on how each screening trial relates the latent *machina* to the probability of selecting the same candidate as *Skynet*. This relationship is best exemplified using item characteristic curves (ICCs), shown in Figure 2, for a selected subset of trials that span different vacancies and identity theme combinations.⁴ In these curves, both the *location* and the *differentiation* parameters can be visually identified. The *location* parameter represents the level of *machina* at which there is a 50/50 probability of matching *Skynet*’s referral. For instance, in Panel A, we plot the relationship between the probability of selecting the same candidate as *Skynet* in both “job vacancy 1”, for which candidates A and B have applied, and in “job vacancy 2”, for which candidates C and D have submitted applications. In this case, we see that these trials have different location parameter levels. In particular, in the latter trial, the cutoff for identifying recruiters (along the latent trait spectrum) who agree more than disagree with *Skynet* is higher (i.e. the dashed curve crosses the 0.5 horizontal marker line to the right of the solid one). Meanwhile, at the point these curves cross the horizontal marker, they have close to identical slopes, signifying identical differentiation parameters. Therefore, these two trials offer contexts that are equivalent in their “latent trait revelation” potential.

In Panel B of Figure 2, we compare “job vacancy 3”, which has E and F as applicants (solid line),

⁴As a reference, the vertical lines indicate the median value of the latent trait among human recruiters.

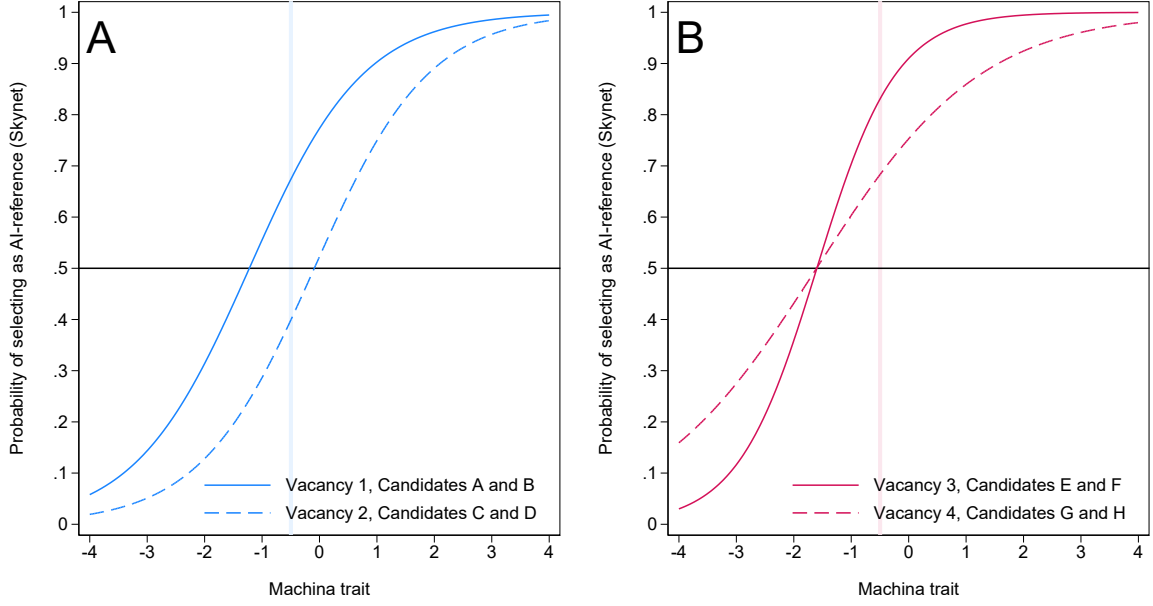


Figure 2: Characterising selected cases of screening trials: trial/item characteristic curves

and “job vacancy 4”, which contrasts candidates G and H (dashed line). Both curves cross the 0.5 probability line at nearly the same point along the *machina* trait, indicating nearly identical location parameters: in these two trials, majoritarian agreement between *Skynet* and recruiters starts to emerge at the same level of *machina* trait. However, their differentiation parameters are distinct: the dashed curve is noticeably flatter, showing that agreement with *Skynet* in the accountant trial is less sensitive to variation in the *machina* trait. This means that the first trial more sharply splits recruiters into groups that are either very machine-like or not, while the latter trial produces a fuzzier differentiation. Taken together with Panel A, these results show that trials differ in how they expose contrasts between human and AI decision-making, either by shifting the threshold for majoritarian agreement with *Skynet* (location) or by how sensitive to differences in the *machina* spectrum they are (differentiation).

Since our main objective is to classify contexts using these parameters, we convert them into a trial-level ordinal metric. We do so by simply locating them by where they lie in the $[0, 1]$ interval captured by their respective cumulative distribution functions (CDF). This transformed variable, therefore, represents the share of trials for which the parameter being studied falls under the respective value for that given trial. For example, *location* parameters with values close to 0 in this ordinal spectrum indicate that in the trial in question, the likelihood of agreement between humans and the reference machine is relatively high for a broader share of humans, while the opposite is true for values near 1. In the case of *differentiation* parameters, the reasoning is similar, with those close to value 1 indicating trials where the distinction (by *machina* trait levels) is relatively sharper between subgroups of recruiters. This ordinal transformation also allows us to easily interpret the subsequent regression exercises as rank regressions which study predictors for parameter (and therefore contextual) heterogeneity.

Location parameters

Motivated by substantive heterogeneity in *location* parameters, we use multivariate ordinary least squares (OLS) regression models to examine systematic relationships between those parameters and the measurable characteristics of the trials, distinguishing across vacancy requirements and differences in applicant attributes (identity and other personal characteristics). We start, however, by simply documenting the average ordinal *location* across job vacancies in Figure 3 below.

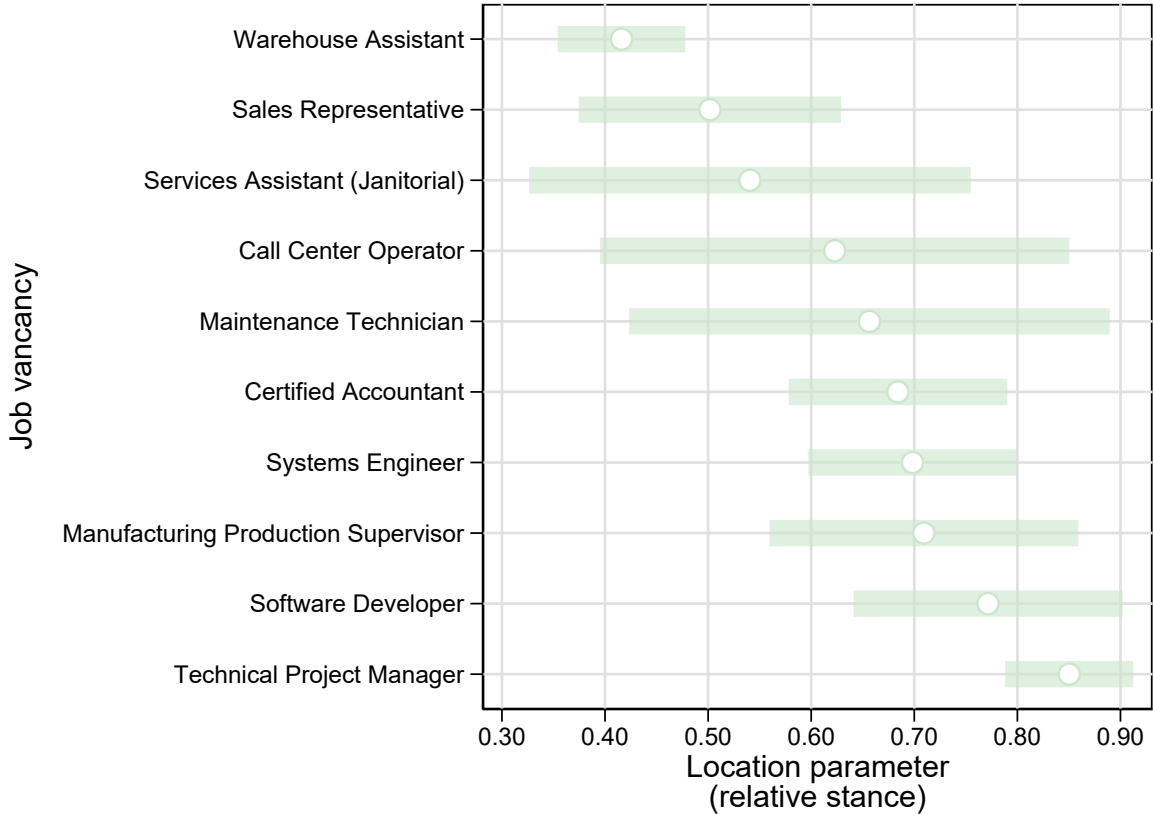


Figure 3: Average *location* (ordinal scale) by trial’s job vacancy type

Looking across specific vacancy descriptions (which implicitly combine education, experience, and specific-skill requirements),⁵ we confirm a gradient across trials’ *location* parameters. Less demanding jobs (in terms of experience, specific skills, and education), such as Warehouse Assistant or Janitorial Services Assistant, exhibit lower *location* rankings, meaning that even less machine-like recruiters tend to relatively often agree with *Skynet* when selecting candidates for these vacancies. More demanding (in terms of skills) positions, such as Technical Manager or Software Developer, have higher *location* rankings, implying that only recruiters with higher values of the latent trait (and, therefore, a smaller share of the human recruiters population align their choices with *Skynet*’s). We conclude that **trials with more sophisticated skill-bundle demands are also the ones in which agreement between human recruiter choices and reference-AI choices is least likely.**

⁵See Supplementary Table S1.

We further confirm this intuition by computing a principal component analysis (PCA) of education and minimum-experience requirements across the 100 trials we designed the field experiment with. We then plot in Figure 4 a simple local bi-variate polynomial rank-rank regression between the *location* parameters (placed in the ordinal scale) and the CDF of the first PCA component (higher values indicating jobs demanding more sophisticated skills).

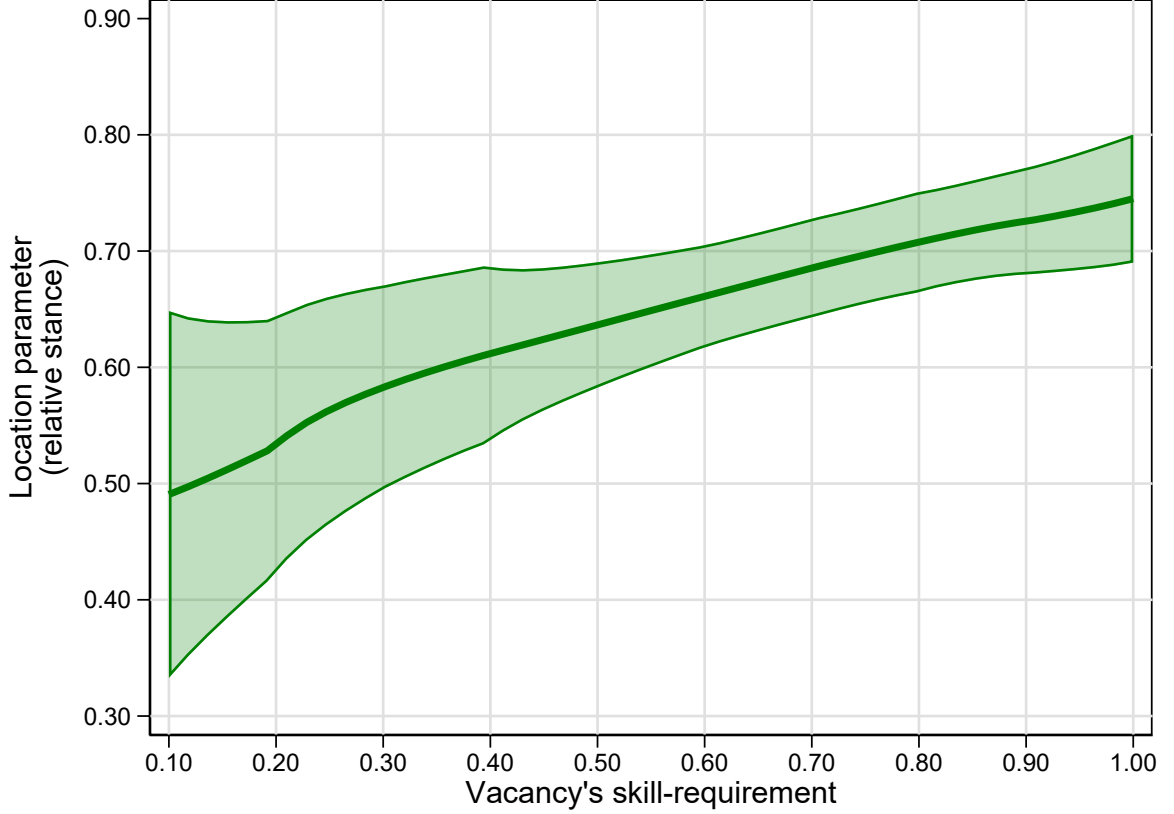


Figure 4: Average *location* by trial/job vacancy’s skill requirements (ordinal scales)

Since each of the trials is also different along candidate-contrast dimensions, we further examine explanatory factors for location parameters. Estimates of the multivariate OLS coefficients explaining heterogeneity in the *location* parameter are displayed in Table 2. We report that the ordinal variation in skill requirement is still a significant predictor of *location* parameter values after we control for the characterisation of candidate contrast along age, experience, and identity dimensions (by construction, schooling is identical for all candidates within CV pairs). A 0.10 increase in the skill requirement CDF corresponds to approximately a 0.04 increase in the *location* parameter CDF. This structure also allows us to employ job-vacancy fixed-effects to focus on the role of CV contrast over the expected level of agreement between human recruiters and *Skynet*. We conclude that **trials with even slightly larger differences in the age or labour experience between candidates are also the ones in which agreement between human recruiter choices and reference-AI choices is least likely to occur**. We hazard that the AI reference-recruiter is particularly attentive to differences that seem irrelevant to the naked human eye, triggering a higher chance of disagreement between them. We also note that there is little systematic difference in *location* across placebo, gender, sexual orientation, and

immigration conditions.

Table 2: *Location* parameter heterogeneity explained by job requirements and candidate differences

	[1]	[2]	[3]
Vacancy’s skill requirement (CDF)	0.250*** (0.086)	0.342*** (0.098)	-
Age difference between candidates (12 mths+)		0.174*** (0.056)	-0.045 (0.137)
Age difference between candidates (6 to 12 mths)		0.191*** (0.066)	0.231** (0.091)
Age difference between candidates (under 6 mths)		REF	REF
Experience difference between candidates (3 to 4 mths)		-0.024 (0.071)	0.279*** (0.062)
Experience difference between candidates (2 to 3 mths)		0.009 (0.076)	0.397*** (0.092)
Experience difference between candidates (under 2 mths)		REF	REF
Gender difference between candidates		0.066 (0.060)	0.053 (0.056)
Immig. status difference between candidates		0.084 (0.076)	0.062 (0.067)
Sexual-orientation difference between candidates		-0.029 (0.059)	-0.033 (0.056)
No identity difference between candidates		REF	REF
Vacancy fixed-effects			YES
Observations	100	100	100
R^2	0.099	0.233	0.445

Differentiation parameters

We conduct the same analysis for *differentiation* parameters in Figure 5. Albeit not identical to what we saw with *location* parameters, we confirm a gradient across trials’ *differentiation* parameters in a similar direction. More demanding jobs (in terms of experience, specific skills, and education), such as Systems Engineer or Certified Accountant, exhibit higher *differentiation* rankings. We also observe this when examining Figure 6.

Our final piece of evidence on the heterogeneity in *differentiation* comes from the multiple regression results presented in Table 3. These estimates confirm the same pattern we saw for *location* parameters. That is to say: both skill demand and differences in age and experience between candidates are associated with trials in which choices are most strongly associated with fluctuations in the *machina* trait.

6.2 *Machina* trait across and within teams

We now turn to a central element of our scale construction, and report on our estimation of the *machina* latent trait. This provides a rich picture of variation across and within teams of recruiters. As mentioned above, in order to aid the interpretation of our findings, we rescale this

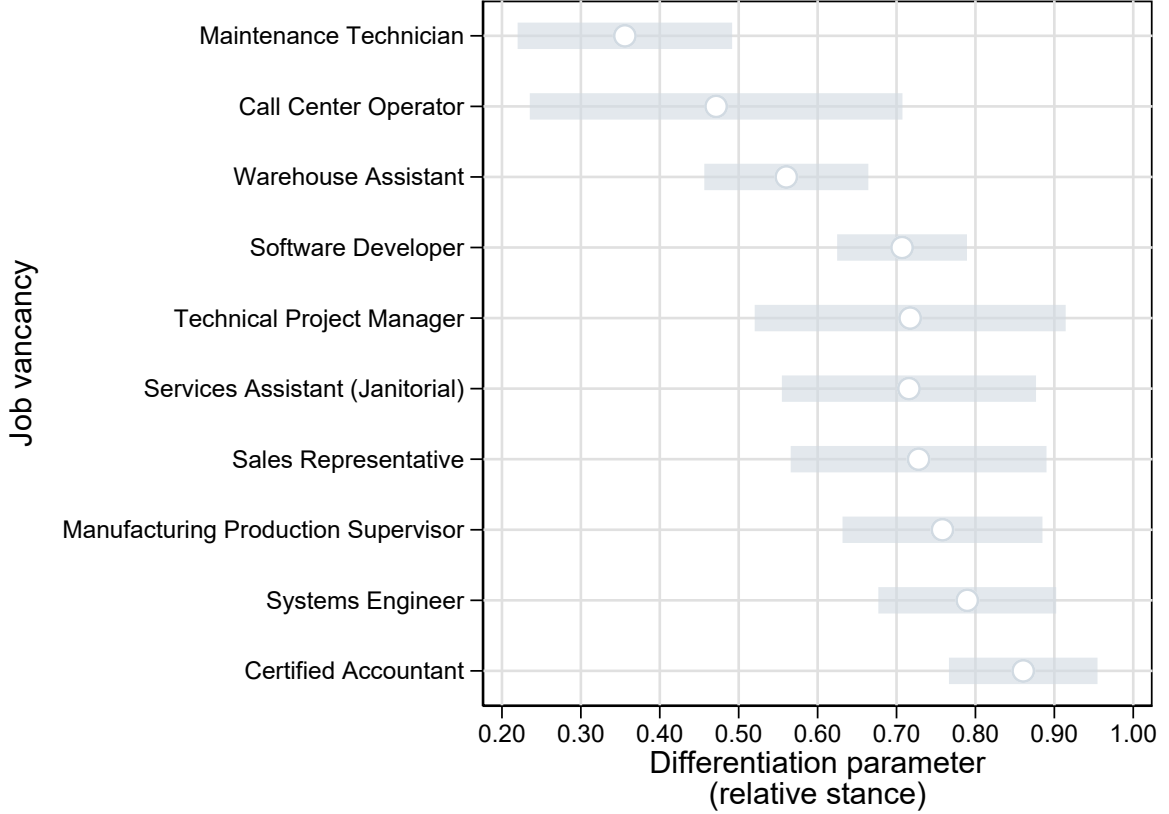


Figure 5: Average *differentiation* (ordinal scale) by trial’s job vacancy type

parameter to be centred around the average human and to be measured in standard deviations within the human recruiters population. We depict the full set of results of our analysis in Figure 7.

We formalise the contrast between these distributions by conducting Kolmogorov-Smirnov tests reported in Table 4, which confirm that no distribution of latent traits among AI recruiting teams is equivalent to the one seen among human recruiters. In sum, as AI agents are endowed with more human-like attributes, their decisions more closely resemble those of human recruiters. However, even the most human-like AI agents (Clones) produced decisions that revealed latent trait distributions that differed significantly from those of human recruiters. The choices made by the Randomistas, who selected candidates at random, endorsed a distribution of *machina* that was closer to that seen amongst humans than those of any of the AI teams.⁶

The bulk of these differences comes from distinct measures of central tendency. We find that the average human recruiter is approximately 1.73 standard deviations less machine-like than the average Robot participating in our experiment (p-value 0.000). The distances to the average Avatar and the average Clone are smaller, 1.51 (p-value 0.000) and 0.80 (p-value 0.000),

⁶This is confirmed by a minimum squared-error fit analysis used for the construction of a mixed distribution between Randomistas and Robots that best approximates the human distribution of latent traits. This results in a best fit that weights by 0.888 the Randomistas density function and by 0.112 the Robots’. Alternatively, a mixed distribution between Randomistas and Clones results in a best fit that weights by 0.720 the former density function and by 0.280 the latter’s.

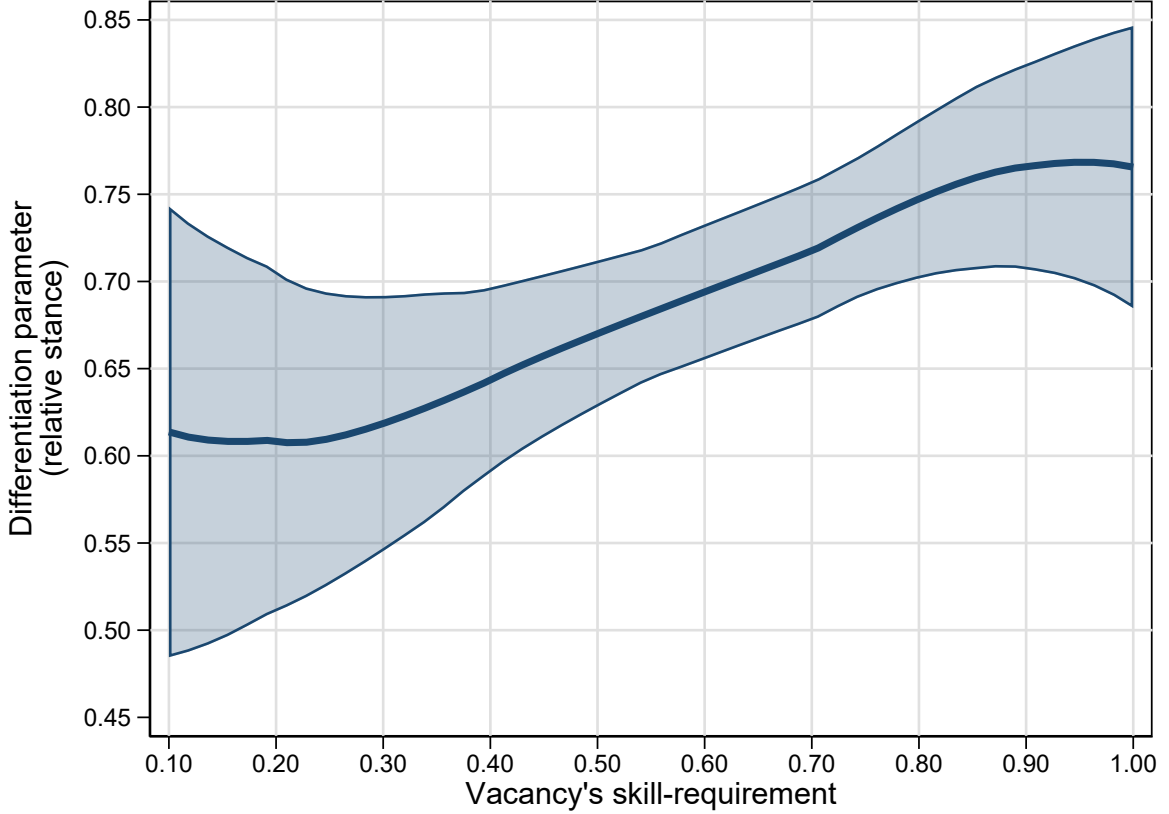


Figure 6: Average *differentiation* by trial/job vacancy’s skill requirements (ordinal scales)

respectively, but still sizable and statistically significant. The average human is also found to be 0.30 standard deviations *closer* to the *Skynet* benchmark choices than the average Randomista. We conclude that while our impersonation and (more so) information restriction instructions approximate the average AI and the average human along the *machina* trait, they do so in an incomplete fashion.

Importantly, these distributions also indicate sizable variation within groups, corroborating the existence of dramatic heterogeneity in the pattern of traits that guide referrals, particularly among human recruiters and AI recruiters subject to both impersonation and information restrictions, as well as among Randomistas. To put these numbers into perspective, we find that the standard deviation in *machina* among humans is equivalent to approximately 58% of the difference in *machina* between the average human and the average Robot. Moreover, this variability is equivalent (according to Levene’s test for homogeneity of variances) to the variability seen among Randomistas and also among Clones, but not among Avatars. This indicates that the incomplete use of job-applicant information may be at the core of the heterogeneity in the latent trait within these recruiter groups we study.⁷

⁷We further delve into this information-centred argument in a companion article (citation withheld for anonymous review).

Table 3: *Differentiation* parameter heterogeneity explained by job requirements and candidate differences

	[1]	[2]	[3]
Vacancy’s skill requirement (CDF)	0.206** (0.099)	0.194** (0.090)	-
Age difference between candidates (12 mths+)		-0.007 (0.067)	0.043 (0.137)
Age difference between candidates (6 to 12 mths)		0.134* (0.078)	0.302*** (0.079)
Age difference between candidates (under 6 mths)		REF	REF
Experience difference between candidates (3 to 4 mths)		0.072 (0.065)	0.135* (0.069)
Experience difference between candidates (2 to 3 mths)		0.140* (0.077)	0.433*** (0.084)
Experience difference between candidates (under 2 mths)		REF	REF
Gender difference between candidates		-0.010 (0.098)	0.016 (0.083)
Immig. status difference between candidates		-0.113 (0.121)	-0.076 (0.096)
Sexual-orientation difference between candidates		-0.021 (0.094)	0.010 (0.080)
No identity difference between candidates		REF	REF
Vacancy fixed-effects			YES
Observations	100	100	100
R^2	0.063	0.221	0.449

6.3 *Machina* trait versus observable human characteristics

Since the human recruiters in our experiment respond to a battery of questions about their professional and personal background, as well as structured modules that allow computation of socio-emotional and cognitive skills, we are also uniquely positioned to assess the extent to which the estimated *machina* can be explained by these observed characteristics. Table 5 presents the results of multivariate least-squares mean and quantile regressions using 277 observations on human recruiters corresponding to the following empirical model:

$$zmachina_i = \alpha_0 + \gamma' \mathbf{x}_i + \epsilon_i, \quad (2)$$

where the vector $\mathbf{x}_i$ captures multiple dimensions of the recruiter’s profiles.

We find that only a few human attributes happen to have a consistent statistically significant relationship with *machina*, however. These include, in particular, neuroticism and self-esteem scores. The latent trait is negatively related to neuroticism, while positively related to self-esteem. We underscore that we do find no relationship between the latent trait and measures of cognitive ability in this model – at least in principle, higher cognitive ability does not make our human recruiter more machine-like.

Importantly, the bulk of the variation on the latent trait among human recruiters cannot be

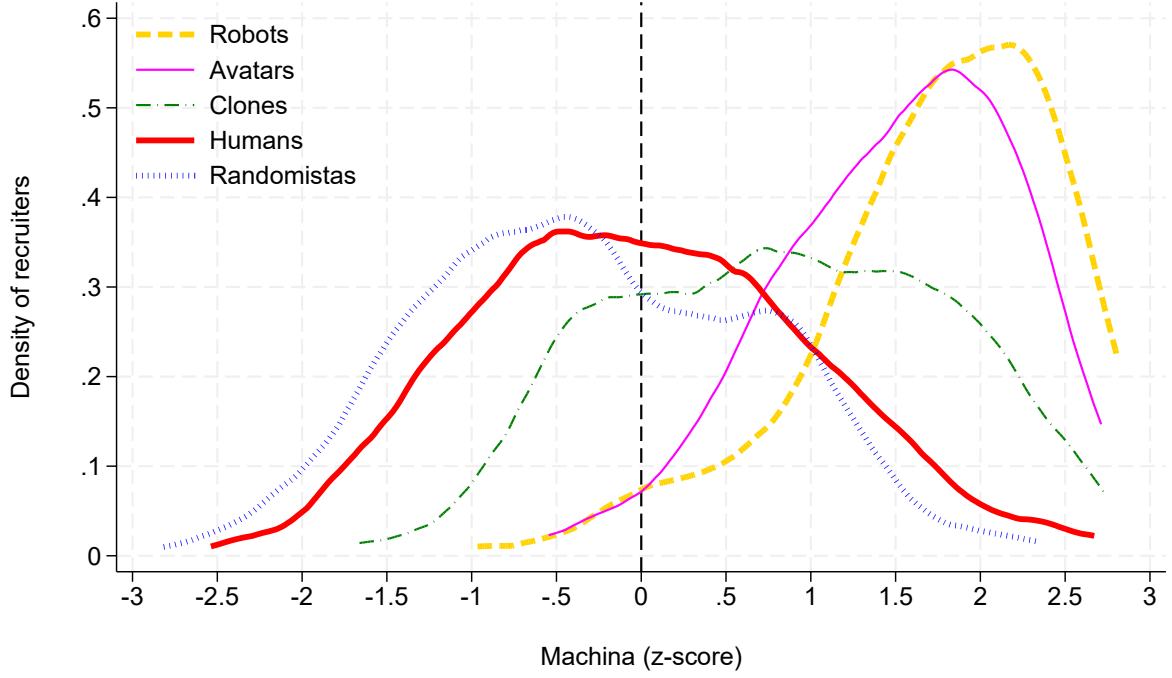


Figure 7: Density of *machina* trait by recruiter team (z-scored based on human recruiters sample moments)

explained by the large set of covariates we observe (R^2 is 0.09). These are important indicators that machine-like choices among human recruiters are poorly explained by a battery of indicators, suggesting that they intrinsically represent a novel behavioural dimension. We, therefore, believe that *machina* can provide more than an analytical metric. It can, for example, offer a lens through which to evaluate the degree to which AI systems conform to rule-based decision norms versus human-like heuristics. As debates around explainable AI and algorithmic accountability evolve, we speculate that carefully designed latent trait models may inform both certification frameworks and governance protocols for high-stakes decision-making systems. We return to this discussion below.

7 Discussion: implications and extensions

Our analysis reveals a fundamental divergence between human and AI decision-making along a common scale. Teams of human recruiters are less systematic in referral choices than their AI counterparts. This heterogeneity in human behaviour is sizable, amounts to more than half of the difference between the average human and the average AI recruiter, and is not explained by a large number of variables captured in our detailed profile of human recruiters (including along cognitive and socio-emotional dimensions). This “behavioural” difference between humans and machines is even more pertinent when vacancies require more credentials (in terms of education and experience) from candidates. Disagreement between machine and humans is also found more likely when the candidates themselves differ in terms of age and labour experience (even if only minutely).

Table 4: Contrasting *machina* trait distribution across recruiter teams

	Mean	SD	Kolmogorov– Smirnov test p -value	Mean equality test p -value	Variance equality test p -value
	[1]	[2]	[3]	[4]	[5]
Robots	1.73	0.72	0.000	0.000	0.000
Avatars	1.51	0.69	0.000	0.000	0.000
Clones	0.80	0.97	0.000	0.000	0.708
Humans	0.00	1.00	REF	REF	REF
Randomistas	-0.30	0.98	0.005	0.000	0.987

Note: Observations are weighted by the precision of the latent trait estimated for each recruiter.

We argue that, implicitly, because observational equivalence was imposed on our experimentally designed candidate pairs, a fair decision rule would assign equal probabilities to each candidate’s being selected. This perspective enables us to reinterpret fairness not merely as the absence of bias, but as the indistinguishability of selection outcomes across equivalent profiles. Interestingly, many human recruiters exhibit behaviours closer to this probabilistic fairness, and they do so more frequently than some AI agents optimised for consistency or efficiency. The latter seems to be hyper-vigilant when it comes to minutia. This framing and findings invite the following reflection: in settings where no observable justification exists for favouring one option over another, true fairness may lie in embracing a degree of randomness.

Our exercises also take steps in the direction of a broader discussion of inconsistencies in human hiring referrals. These are invariably found to be prone to bias, particularly when individuals from marginalised groups are being evaluated, such as those defined by gender, immigration status, or sexual orientation, as shown in labour market discrimination studies (Bartlett et al., 2022; Dressel and Farid, 2018; Fabregas and Zanoni, 2025; Zanoni et al., 2024a,b). These limitations can be understood through the lens of bounded rationality, where cognitive constraints lead decision-makers to fail to optimise (Simon, 1955). Recent work has begun to extend this framework to human–AI interaction itself, examining how trust in algorithmic outputs is shaped by the same constraints on attention, learning, and persuasion that govern human reasoning (Almutairi and Almatrodi, 2025). AI systems, with their ability to process large datasets, may move closer to the format of rational “boundlessness” (Shick et al., 2024); however, they still face some constrained capacity concerning processing information, mainly due to finite computing power and incomplete or biased training data (Jirásek, 2024; Ma and Su, 2024). We point out that in our exercises above, identity themes played no role on the expected level of agreement between (wo)men and machine, nor on the sensitivity of this agreement level with respect to the latent trait we construct.

Yet, if AI recruiters can deliver more consistent and impartial assessments, they hold the potential to reduce inequality and improve hiring efficiency. However, when AI systems are trained on human decisions or designed to imitate human behaviour (Pan et al., 2020; Hagendorff et al., 2023) or when they operate with limited information on productivity-relevant attributes, they risk replicating existing biases under the appearance of objectivity. Closely related evidence

Table 5: *Machina* trait versus covariates: mean and median regressions

	Mean Regression (precision weights) [1]	Mean Regression (unweighted) [2]	Median Regression [3]
<i>Demographics and Education</i>			
Male (=1)	−0.053 (0.128)	−0.070 (0.132)	0.066 (0.152)
Age (years)	−0.017 (0.011)	−0.016 (0.011)	−0.023 (0.016)
Postgraduate degree (=1)	0.320 (0.233)	0.318 (0.242)	0.472 (0.304)
College degree (=1)	−0.036 (0.209)	−0.016 (0.216)	0.032 (0.303)
HR-related major (=1)	0.093 (0.160)	0.077 (0.168)	0.125 (0.214)
<i>Work experience</i>			
Years of HR experience	0.019 (0.030)	0.022 (0.032)	0.047 (0.034)
Currently employed (=1)	0.027 (0.125)	0.034 (0.129)	−0.022 (0.183)
HR-related current/last job (=1)	−0.228 (0.224)	−0.250 (0.227)	−0.432* (0.256)
HR experience abroad (=1)	0.011 (0.177)	−0.006 (0.186)	−0.005 (0.219)
High knowledge of Ecuadorian market (=1)	−0.007 (0.153)	0.025 (0.156)	−0.058 (0.209)
<i>Socio-emotional and Cognitive Scores</i>			
Neuroticism (z-score)	−0.151* (0.079)	−0.156* (0.079)	−0.232** (0.094)
Extroversion (z-score)	0.017 (0.086)	−0.002 (0.090)	0.004 (0.104)
Openness (z-score)	−0.039 (0.071)	−0.032 (0.072)	0.007 (0.084)
Agreeableness (z-score)	0.039 (0.088)	0.045 (0.091)	0.078 (0.109)
Conscientiousness (z-score)	−0.039 (0.080)	−0.044 (0.082)	−0.021 (0.094)
Self-esteem (z-score)	0.131** (0.061)	0.134** (0.065)	0.202** (0.091)
Cognitive test (z-score)	0.029 (0.064)	0.034 (0.066)	0.110 (0.083)
Observations	277	277	277
R^2	0.093	0.093	
Pseudo- R^2			0.089

Note: Infinitesimal-jackknife (Eicker-White) standard errors in parentheses under point estimates for mean regressions. All models include indicators for the sector of professional experience and an indicator for recruiter hiring approach (LinkedIn or snowball referral). Observations are weighted by the precision of the latent trait estimated for each recruiter for mean regressions. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

comes from public-opinion simulation experiments showing that the fidelity of LLM outputs to human response distributions varies systematically with cultural and demographic context and degrades as the choice set becomes more complex (Qu and Wang, 2024), which is a pattern consistent with our finding that human-AI agreement deteriorates in trials with more demanding skill requirements and more subtle candidate contrasts. Moreover, algorithmic decisions may

lack legitimacy if perceived as detached from human values or fairness norms (Lee and Baykal, 2017; Danaher, 2016), and human trust in AI varies due to algorithm aversion or appreciation (Dietvorst et al., 2015; Logg et al., 2019), with recent reviews emphasising that trust in AI is shaped by the perceived alignment between algorithmic outputs and human values (Bach et al., 2024). Understanding how AI converges with or diverges from human behaviour, particularly the making of choices in the presence of identity-based cues, is thus essential for evaluating the former’s role in hiring. The observed differences in decision-making between human and AI recruiters are particularly salient, given that our AI agents were not trained on any human decision-making data. Their behaviour, therefore, stems from their underlying pre-trained models and the specific prompts designed to imbue them with human-like traits or contextual information, rather than from the explicit learning of human biases or inconsistencies from an empirical dataset. In this sense, AI does not produce pure objectivity; rather, it renders a more stable, rule-based intersubjective reality, one that mirrors societal preferences rather than transcending them.

Nonetheless, screening to inform about a candidate’s unobserved productivity requires interpreting complex, multidimensional, and often noisy signals. Both humans and AI rely on heuristics to navigate this uncertainty (Stiglitz, 1975). As a result, the divergence in their decisions arises not only from differences in bounded rationality but also from how each processes incomplete and imperfect indicators of human potential. Ultimately, both human and AI recruiters operate under the dual constraints of cognitive limitations and informational gaps, underscoring the inherent complexity of hiring decisions. Resorting to AI to address these challenges may appear promising, but doing so without fully understanding how algorithmic logic differs from human reasoning risks introducing new distortions rather than resolving existing ones. Our results document that the differences are substantial and are also affected by context (even within a simplified, yet realistic, candidate screening exercise).

Our findings also suggest a core paradox in AI-driven recruitment: the more “AI-like” a recruiter becomes—exhibiting consistency, rule-based reasoning, and alignment with an idealised benchmark—the less human its decision-making appears. This divergence reduces interpretability and relatability from a human perspective. Clones, designed to impersonate individual human recruiters, illustrate this tension: as AI agents become more human-like, they introduce greater variability, possibly undermining the consistency that makes algorithmic systems attractive in the first place.

Our Labour Market Turing Test (LMTT) and cognitive interviewing approach offer a potentially generalisable framework for evaluating AI–human alignment with an eye to the ultimate purpose of ensuring algorithms are consistent, fair, and trustworthy across fields. Our methods and findings may also contribute to the broader (and more normative) AI research agenda of constructing a *machina economicus*, synthetic rational agents capable of operating effectively in multi-agent economic environments (Parkes and Wellman, 2015). While our analysis focuses on hiring, the implications extend far beyond this context. Similar challenges emerge across a range of high-stakes domains where AI systems are increasingly being tasked with replicating, pairing with, or replacing the *homo economicus*.

References

- Almutairi W, Almatrodi I (2025) Trust under bounded rationality: exploring human–AI interaction in decision-making through large language models. *SAGE Open* 15(4):21582440251380135. <https://doi.org/10.1177/21582440251380135>
- Amershi S, Weld D, Vorvoreanu M, Fournery A, Nushi B, Collisson P, Suh J, Iqbal S, Bennett PN, Inkpen K, Teevan J, Kikin-Gil R, Horvitz E (2019) Guidelines for human-AI interaction. In: Brewster S, Fitzpatrick G (eds) *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, pp 1–13
- Bach TA, Khan A, Hallock H, Beltrão G, Sousa S (2024) A systematic literature review of user trust in AI-enabled systems: an HCI perspective. *Humanities and Social Sciences Communications* 11:1568. <https://doi.org/10.1057/s41599-024-04044-8>
- Bartlett R, Morse A, Stanton R, Wallace N (2022) Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics* 143(1):30–56
- Birnbaum A (1968) Some latent trait models and their use in inferring an examinee’s ability. In: Lord FM, Novick MR (eds) *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, pp 397–479
- Chen Z (2023) Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications* 10(1):567. <https://doi.org/10.1057/s41599-023-02079-x>
- Danaher J (2016) The threat of algocracy: reality, resistance and accommodation. *Philosophy and Technology* 29(3):245–268. <https://doi.org/10.1007/s13347-015-0211-1>
- Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114–126
- Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4(1):eaao5580
- Eriksson T, Johansson P, Langenskiöld S (2017) What is the right profile for getting a job? A stated choice experiment of the recruitment process. *Empirical Economics* 53(2):803–826
- Fabregas R, Zandoni W (2025) The migrant penalty in Latin America: experimental evidence from job recruiters. *Journal of Public Economics* 247:105393. <https://doi.org/10.1016/j.jpubeco.2025.105393>
- Hagendorff T, Fabi S, Kosinski M (2023) Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3(10):833–838
- Hoffman M, Kahn LB, Li D (2018) Discretion in hiring. *Quarterly Journal of Economics* 133(2):765–800

- Jirásek M (2024) DAOs: a lens on AI and bounded rationality. SSRN Working Paper No. 4703706
- John OP, Donahue EM, Kentle RL (1991) The Big Five Inventory: versions 4a and 54. Institute of Personality and Social Research, University of California, Berkeley
- Kaufman M (2025) AI might be scanning your resume. Here’s what job hunters should know. CNN, 8 April. <https://www.cnn.com/2025/04/08/tech/ai-resume-job-hunters>
- Lee MK, Baykal S (2017) Algorithmic mediation in group decisions: fairness perceptions of algorithmically mediated vs. discussion-based social division. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp 1035–1048. <https://doi.org/10.1145/2998181.2998230>
- Li D, Raymond L, Bergman P (2020) Hiring as exploration. NBER Working Paper No. 27736. Forthcoming, Review of Economic Studies
- Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, others (2019) A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* 1(6):e271–e297
- Liu SH, Juster RP, Dams-O'Connor K, Spicer J (2020) Allostatic load scoring using item response theory. *Comprehensive Psychoneuroendocrinology* 3:100009. <https://doi.org/10.1016/j.cpnec.2020.100009>
- Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: people prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Lord FM, Novick MR (1968) Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA
- Ma H, Su M (2024) The bounded intelligence of AI: superficiality and deceivability. *Organizational Dynamics* 53:101100
- May H (2006) A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education. *International Journal of Testing* 6(1):1–24. https://doi.org/10.1207/s15327574ijt0601_1
- Microsoft (2025) Prompt engineering concepts in semantic kernel. <https://learn.microsoft.com/en-us/semantic-kernel/concepts/prompts>. Accessed 15 July 2025
- Pan M, Huang W, Li Y, Zhou X, Luo J (2020) xGAIL: explainable generative adversarial imitation learning for explainable human decision analysis. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 1334–1343
- Parkes DC, Wellman MP (2015) Economic reasoning and artificial intelligence. *Science* 349(6245):267–272. <https://doi.org/10.1126/science.aaa8403>

- Qu Y, Wang J (2024) Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications* 11(1):1095. <https://doi.org/10.1057/s41599-024-03609-x>
- Rosenberg M (1965) *Society and the Adolescent Self-Image*. Princeton University Press, Princeton
- Shaghaghi A, Bhopal RS, Sheikh A (2011) Approaches to recruiting ‘hard-to-reach’ populations into research: a review of the literature. *Health Promotion Perspectives* 1(2):86–94. <https://doi.org/10.5681/hpp.2011.009>
- Shen Y, Zhang X (2024) The impact of artificial intelligence on employment: the role of virtual agglomeration. *Humanities and Social Sciences Communications* 11(1):122. <https://doi.org/10.1057/s41599-024-02647-9>
- Shick M, Johnson N, Fan Y (2024) Artificial intelligence and the end of bounded rationality: a new era in organizational decision making. *Development and Learning in Organizations* 38(4):1–3
- Simon HA (1955) A behavioral model of rational choice. *Quarterly Journal of Economics* 69(1):99–118
- Stiglitz JE (1975) The theory of ‘screening’, education, and the distribution of income. *American Economic Review* 65(3):283–300
- Turing AM (1950) Computing machinery and intelligence. *Mind* 59(236):433–460
- Turner A (2025) The psychology of prompt engineering: understanding user interaction with AI. <https://arsturn.com/blog/the-psychology-of-prompt-engineering-understanding-user-interaction-with-ai>. Accessed 15 July 2025
- Vanderbilt University Generative AI Initiative (2025) Prompt patterns. <https://www.vanderbilt.edu/generative-ai/prompt-patterns>. Accessed 15 July 2025
- Will P, Krpan D, Lordan G (2023) People versus machines: introducing the HIRE framework. *Artificial Intelligence Review* 56(2):1071–1100
- Willis GB (2005) *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. SAGE Publications, Thousand Oaks, CA
- Wiwad D, Mercier B, Maraun MD, Robinson AR, Piff PK, Aknin LB, Shariff AF (2019) The support for economic inequality scale: development and adjudication. *PLOS ONE* 14(6):e0218685. <https://doi.org/10.1371/journal.pone.0218685>
- Wonderlic EF (1999) *Wonderlic Personnel Test Manual*. Wonderlic Inc., Libertyville, IL
- Zanoni W, Duryea S, Paredes J (2024a) Gender and discrimination in urban labor markets: evidence from a field experiment in Ecuador. SSRN Electronic Journal. <https://ssrn.com/abstract=5068950>

Zanoni W, Hernández H, Zambrano O, Quiroz G (2024b) Discrimination against lesbian and gay job seekers: an artefactual field experiment in urban Ecuador. *Labour Economics* 88:102531. <https://doi.org/10.1016/j.labeco.2024.102531>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SIDevelopingAndAssessingAug2026.pdf](#)