WORKING PAPER Nº IDB-WP-01740

Documenting Differences Between Humans and AI in High-Stakes Decisions: A Labor Market Turing Test

Andres Abril
Marcos A. Rangel
Wladimir Zanoni

Inter-American Development Bank
Country Department Andean Group

September 2025

Documenting Differences Between Humans and AI in High-Stakes Decisions: A Labor Market Turing Test

Andres Abril[1]
Marcos A. Rangel[2*]
Wladimir Zanoni[3**]

[1]Universidad de las Americas
[2] Duke University
[3]Inter-American Development Bank

Inter-American Development Bank
Country Department Andean Group

September 2025

Cataloging-in-Publication data provided by the

Inter-American Development Bank

Felipe Herrera Library

Abril, Andres.

Documenting differences between humans and AI in high-stakes decisions: a labor market Turing test / Andres Abril, Marcos A. Rangel, Wladimir Zanoni.

p. cm. — (IDB Working Paper Series ; 1740)

Includes bibliographical references.

1. Turing test. 2. Artificial intelligence-Ecuador. 3. Labor supply-Ecuador. 4. Employee selection-Ecuador. I. Rangel, Marcos. II. Zanoni, Wladimir. III. Inter-American Development Bank. Country Department Andean Group. IV. Title. V. Series.

IDB-WP-1740

Keywords: Algorithmic Fairness, Human-AI Alignment, Latent Trait Analysis.

JEL Codes: J71, M51, C91.

http://www.iadb.org

Rea_cam@iadb.org

Documenting Differences Between Humans and AI in High-Stakes Decisions: A Labor Market Turing Test

Andres Abril,[1] Marcos A. Rangel,[2*] Wladimir Zanoni[3**]

[1]Universidad de las Americas
[2] Duke University
[3]Inter-American Development Bank
$*$ marcos.rangel@duke.edu,$**$ wladimirz@iadb.org

We introduce a Labor Market Turing Test (LMTT) to evaluate alignment between human and AI decision-making in hiring. Through a randomized lab-in-the-field experiment with human recruiters, we compare their referral decisions to those of AI-recruiter teams built on large language models that sequentially impersonate traits of human counterparts. We define *machina*, a latent trait mapping human and AI recruiting decisions along a common scale. Results reveal distributions of such traits among humans differ distinctly from AI-based ones across first and higher moments. While impersonation of human traits approximates AI teams' variance and skewness to human levels, full capture remains elusive. Human patterns are better mapped through mixed distributions heavily weighting random selection over AI-rational candidate choices. Importantly, trait alignment between human and AI decisions shifts systematically with job complexity and minor applicant distinctions. The LMTT thus offers both a general measurement method and new insights into human-AI decision alignment trade-offs.

Introduction

Screening and selection processes influence nearly every aspect of modern life, from our personal relationships to key institutional decisions in hiring, healthcare, criminal justice, and education. Once driven exclusively by human judgment, these decisions are now increasingly guided—or even entirely made—by artificial intelligence (AI). The portrayal of AI as an objective decision-maker is often a misconception, as it conflates the systematic nature of an algorithm with true impartiality. The central research question is not whether AI can replicate human decision-making, but rather to critically examine how its systematic processes may reproduce and reinforce human biases against specific societal groups.

The screening processes within labor markets provide a particularly compelling case study for understanding the divergence between human and algorithmic decision-making. 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. With some job postings attracting hundreds or even thousands of applications, companies increasingly rely on AI to screen applicants before human recruiters review them. According to recent press coverage (*1*), 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 utilization through 2030. Given the significant impact hiring decisions have on individuals' lives and the increasing complexity of labor markets, AI is poised to become a major gatekeeper in mediating access to economic opportunities. This transformation has far-reaching implications for both efficiency and equity.

In this paper, we examine how job referral decisions differ within and between teams of human and AI recruiters while documenting which aspects of their decision-making processes

drive these differences. While prior research has compared algorithmic versus human performance at screening (*2–4*), few studies have implemented controlled experiments to directly measure alignment between human and AI judgments or to identify the conditions under which such alignment is more or less likely to break down. Using data from a lab-in-the-field randomized control trial with real recruiters in Ecuador, we evaluate how AI recruiter teams perform the same candidate selection tasks relative to their human counterparts. These AI teams were constructed to vary in the degree to which they are asked to impersonate humanlike traits, as defined by the profiles captured by survey instruments filled by human recruiters. Specifically, we measure differences in decisions when recruiters (human and AI) assess observationally identical pairs of job applicants, with one applicant randomly assigned to a minority group (migrant, woman, or self-identified as gay or lesbian).

To systematically compare the human and AI decisions, we develop a Labor Market Turing Test (LMTT) framework. Leveraging psychometric techniques, we define a latent trait, *machina*, that captures how AI-like an agent's choices (human or otherwise) appear. The method generates a unidimensional metric of behavioral distance between each AI agent and each human recruiter in our experiment. Ultimately, it allows us to compare the decisions and behavioral 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 technique reveals how particular job specifications and candidate profiles help distinguish humans from machines, clarifying how context shapes these contrasts.

Experimental evidence on recruiting by humans

To evaluate hiring referral decisions, we draw on data generated by 277 freelance human resources recruiters we hired for a lab-in-the-field experiment conducted in Ecuador. This experiment, which we described in detail in the Supplementary Materials (SM, Section SM1), was

previously used to examine patterns of unequal treatment (as a function of applicant's attributes) in hiring decisions (*5–7*). Using a web-based platform mimicking standard human resources software (see Figures SM3 and SM4), the recruiters were tasked with referring best-fit candidates for each of 10 job vacancies with which they were presented, each one with exactly two applicants (Section SM2). 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 job vacancies, each had candidate pairings randomly assigned (out of 10 alternative pairings per vacancy), based on an experimental structure we describe diagrammatically in Figure SM5. We henceforth refer to each of these decisions as trials. Ultimately, each recruiter was exposed to a unique sequence of 10 vacancy-candidate pair trials, with each trial being evaluated by 17 to 40 unique human recruiters by the end. Within each trial, job candidates were roughly matched on productivity-related traits (e.g., education, experience, age) and differed only in one randomized identity marker theme, either gender, Venezuelan origin (migrant), or sexual orientation, as detailed in Section SM3. The latter should, in principle, be irrelevant to candidate selection because they do not explicitly represent productive attributes among the jobs we were hiring for.

Recruiters also completed a detailed background survey, including demographic characteristics, professional experience, and performance on standardized cognitive, socio-emotional, and labor market knowledge tests (See Table SM2). Recruiters in our sample (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), and 1 in 5 had completed some graduate-level program. Participants had, on average, 3.1 years of work experience, mostly in HR-related roles. Approximately 61% were employed at the time of participation, and 13% had some work experience outside Ecuador. Responses to labor market knowledge questions (regarding minimum wage and basic labor rights, for example)

indicated strong awareness, with 86% scoring in the top category. In terms of socio-emotional traits, we observe that our average 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, surprisingly, may not be well-suited to jobs involving advanced reasoning and learning.

AI recruiters

We developed four AI recruiter teams using ChatGPT (version GPT 4o-mini), each comprising 277 agents, mirroring the structure and sample size of the human recruiter sample. To guide the development of the AI recruiters, we adapted the method of *cognitive interviewing* (*8*)—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 behavior through structured interactions to ensure comprehension and clarify ambiguity. This technique is consistent with Human-AI Interaction (HAIC) principles that emphasize trust, interpretability, and adaptability in AI systems (*9–12*). Through structured dialogue, we assessed the AI's understanding of the task, its interpretation of humanlike traits they 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 behavior in high-stakes decision contexts such as hiring. It also facilitates a conceptual bridge between human cognition and algorithmic reasoning, one of the core inquiries of our research agenda.

Figure SM6 is a summary infographic explaining the creation of the AI recruiter teams. All our AI agents followed a standardized 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 behavior, we relied on their inherent, pre-existing algorithmic knowledge. The four AI teams differed only in the degree of instruction they received before executing the screening task. Each team was provided access to different types of contextual information, organized 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).

The teams were defined by the contextual information blocks to which they were subject. We adopted the following labeling: *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 realization of a coin toss every time a choice was required. This gradation from clones to randomistas illustrates a range of possible alignments with human context and behavior. More details are presented in Section SM4.

A Labor Market Turing Test

In order to systematically contrast human and algorithmic judgments under these conditions, we propose a variant of the classic Turing Test (*13*), adapted to personnel hiring: a *Labor Market Turing Test* (LMTT). The LMTT formalizes 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 conceptualization is based on the premise that while both humans and machines must navigate similar constraints, they likely do so in distinctive manners: humans draw on tacit knowledge and socio-emotional reasoning, while AI leverages 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 (*14*) and paired comparison protocols from machine learning (*15*). 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 candidate pairs and asked to select exactly one as the most qualified for the job (Figure SM7).

The binary structure of the screening tasks also aids the multidimensional contrast across multiple recruiters we set out to develop here. We propose a model-based framework that places all decision-makers—human or AI—on a common evaluative scale. Specifically, we draw from a long tradition in psychometrics and use a two-parameter item response theory (IRT) model (*16, 17*). 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 and physiological dysregulation (*18, 19*), 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 full sequence of candidates being evaluated by a given recruiter is different from the one evaluated by the next.

More specifically, our IRT model assumes that there exists an unobservable latent (and con-

ceptual) trait, which we call *machina*, that cannot be directly measured, unlike observable characteristics such as human cognition 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 behavior of a representative AI agent. The choices of the representative AI agent serve as a conceptual answer key, providing a benchmark for what a perfectly machinelike 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 (or central planner) as *Skynet*, conceived to be representative of pure machinelike consensus behavior which we fully describe in Section SM5. It is crucial to distinguish between *Skynet*, which serves as a benchmark for prototypical machinelike behavior reflecting inherent algorithmic priorities, and the concept of fairness in high-stakes decisions. This is because while *Skynet* provides a measure of pure algorithmic consistency, it does not, by itself, represent a normative ideal or an inherently fair outcome.

The IRT estimation procedure involves an expectation-maximization (EM) iterative algorithm solution. 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.

In addition to this full characterization of heterogeneity across recruiters in their propensity to agree with the reference machine, 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. This is analogous to the "difficulty" parameter in IRT, which measures at what level of proficiency test takers start having more than 50/50 chance to get the right answer in a given question. Here, however, our location parameter places a given trial on the *machina* trait continuum, indicating the trait level at which recruiter are equally

likely to agree or disagree with *Skynet*'s choice. 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. Therefore, our modeling exercise can reveal screening trials that span a broad spectrum of predictability levels—some associated with high *machina*, others with moderate *machina*, and some with low *machina*.

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. 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 machinelike from the non-machinelike behavior.

Empirical findings

We estimate our model using a Newton-Raphson numerical method and employ the binary choices of each member of the team of 277 human recruiters and the aforementioned four teams of AI recruiters (each also containing 277 members).

Machina trait

Our estimation of the *machina* latent trait provides a rich picture of variation across and within teams of recruiters. To aid the interpretation of our findings, we rescale this parameter to be centered around the average human and to be measured in standard deviations within the human

recruiter population. We depict the full set of results of our analysis in Figure 1. We find that the average human recruiter is approximately 1.77 standard deviations less machinelike than the average Robot participating in our experiment (p-value 0.000). The distances to the average Avatar and the average Clone are smaller, 1.53 (p-value 0.000) and 0.80 (p-value 0.000), 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.

We also detect sizable variation within groups, indicating important heterogeneity in the pattern of 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 57% 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, indicating 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.

We formalize the contrast between these distributions by conducting Kolmogorov-Smirnov tests, 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 humanlike attributes, their decisions more closely resemble those of human recruiters. However, even the most humanlike AI agents (Clones) produced decision 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 AI team. This 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 0.891 the Randomistas density function and by 0.109 the Robots'.

Because the human recruiters 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 can also assess the extent to which the estimated *machina* can be explained by these observed characteristics. Table SM4 presents the results of multivariate mean and median regressions using 277 observations on human recruiters. There are a few human attributes that happen to have a statistically significant relationship with *machina*, however. These include age and holding a post-graduate degree, as well as neuroticism and self-esteem scores. Age is negatively related to, while those with more educational credentials seem to score higher on, the *machina* trait. Interestingly, the latent trait is also negatively related to neuroticism, while positively related to self-esteem. Quantile regressions at the bottom and top third of the distribution also reveal that the dispersion of the latent trait increases across these same four dimensions, indicating an interesting pattern of heteroskedasticity in the *machina* trait among human recruiters

Yet the bulk of the variation on the latent trait among human recruiters cannot be explained by the covariates we observe (R-squared is 0.09). These are important indicators that machine-like choices among human recruiters are poorly explained by a battery of covariates, indicating that they intrinsically represent a novel behavioral dimension. We, therefore, believe that the *machina* provides more than an analytical metric—it offers a lens through which to evaluate the degree to which AI systems conform to rule-based decision norms versus humanlike heuristics. As debates around explainable AI and algorithmic accountability evolve, carefully designed latent traits models may inform both certification frameworks and governance protocols for

high-stakes decision-making systems.

Trial-level contextual characterization

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 subset of trials that span different vacancies and identity theme combinations. In these curves, both the "location" parameter and the "differentiation" parameter 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 the sales representative vacancy in which no candidate identity theme was highlighted (placebo) and in the technical project manager position in which sexual orientation was different between candidates. 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 dotted 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.

In Panel B of Figure 2, we compare a warehouse assistant vacancy with male and female applicants (solid line) and an accountant vacancy with immigrant versus non-immigrant applicants (dotted line). Both curves cross the 0.5 probability line at nearly the same point along the *machina* trait, indicating similar location parameters: in these two trials majoritarian agreement between *Skynet* and recruiters start to emerge at the same level of *machina* trait. However, their differentiation parameters is distinct: the dotted curve is noticeably flatter, showing that agreement with *Skynet* in the accountant trial is less sensitive to variation in the *machina* trait. This

means the warehouse assistant vacancy more sharply splits recruiters into groups that either more machinelike and those who are less machinelike, while the accountant vacancy 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 changing how sharply agreement varies along the *machina* spectrum (differentiation).

Motivated by this descriptive evidence, we next use bi-variate ordinary least squares (OLS) regression models to examine systematic relationships between these parameters and the characteristics of the trials, distinguishing across applicant attributes (the randomized identity theme, and personal characteristics), and vacancy requirements (see descriptives in Table SM3 and specification details in Section SM7). Estimates of the OLS coefficients relating the *location* and *differentiation* with vacancy requirements are presented in Figures 3 and 4, respectively. Panel A in Figure 3 shows that vacancies requiring a college degree exhibit significantly higher *location* values, meaning alignment with *Skynet* more restricted to recruiters in the upper part of the *machina* distribution for those. The vertical lines place reference values for these parameters using percentiles of the latent trait among human recruiters. Vacancies that require a college degree, for example, see agreement with *Skynet* crosses the 50/50 benchmark at a level of the trait that correspond to the first quartile of the human *machina* trait distribution. This is in contrast with vacancies mot requiring college, with location parameters lower than the 10th percentile of the same distribution. Panel B in 3 shows that more experience requirements also correspond to higher *location*, but the differences are less pronounced than for the case of education.

Looking across vacancy types (which implicitly combine education and experience requirements) in panel C of Figure 3, we confirm the gradients across these two dimensions. Less demanding jobs, such as Warehouse Assistant or Janitorial Assistant, exhibit lower *location* estimates, meaning that even less machinelike recruiters tend to agree with *Skynet*. More de-

manding positions, such as Technical Manager or Software Developer, have higher *location* estimates, implying that only recruiters with higher values of the latent trait align with *Skynet*.

This contrasts follow a similar pattern for the case of *differentiation* parameters. Panel A in Figure 4 shows that trials where the vacancy pertains to a college-required job display higher sharpness in differentiating recruiters who align with *Skynet* from those who do not. When it comes to differences in experience required, panel B of the same figure indicates a flatter slope, suggesting less ability to differentiate agreement with *Skynet*. In Panel C, we also see that jobs such as Call Center Operator and Maintenance Technician exhibit relatively low values, while Accountant and System Engineer positions yield some of the steepest slopes, making of them trials where the separation between more and less machine-like recruiters is particularly sharp.

In Figure 5, we plot the coefficient estimates of the indicated relationships. Starting with identity contrasts (panels A and B), we find little systematic difference in *location* across placebo, gender, sexual orientation, and immigration conditions. However, we can highlight a tendency toward higher *location* parameters in immigrant and gender trials, implying that more machine-like recruiters are required for agreement with *Skynet*. Along the *differentiation* dimension, identity differences seem to matter more importantly: immigrant contrasts yield flatter slopes, indicating weaker separation between recruiters by their latent trait, while the other identity themes display more sensitivity to variation in *machina*.

Turning to experience gaps between applicants in panels C and D, those trials with small differences in experience (up to two months) show almost no difference along the *location* parameter relative to those portraying more than three months of difference in experience. However, when we turn attention to the *differentiation* parameter, it becomes noticeable that larger experience gaps are associated with more differentiation, sharpening the distinction between recruiters who align with *Skynet* and those who do not. In the case of age differences, panels E and F in Figure 5 show that those trials with candidates more than twelve months apart show

higher *location* parameters, indicating that agreement with *Skynet* occurs only among more machine-like recruiters in these comparisons. Heterogeneity in the *differentiation* values by age differences is practically inexistent, suggesting that age-related contrasts create little distinction between recruiters in their alignment with the latent *machina* trait.

Formal statistical tests of these differences, reported in Table SM5, confirm the strong and significant effect of education: college requirements consistently raise *location* parameters. Larger age gaps between candidates also increase *location* estimates, with weaker evidence of effects from identity contrasts. Stratified models show additional heterogeneity: among jobs without a college requirement, both identity and age differences drive higher *location* values, while in college-required jobs, wider experience gaps lower *location* values. *Differentiation* patterns also vary: higher experience requirements reduce *differentiation* only in jobs without a college requirement, whereas gender and experience gaps increase *differentiation* in jobs with a college requirement. Among jobs not requiring college, only age differences appear positively related to *differentiation*.

Discussion: Implications and extensions

Our analysis reveals a fundamental divergence between human and AI decision-making. Human recruiters are, on average, less systematic than their AI counterparts, and even the most humanlike AI agents produce choices that remain distinct from human patterns. AI's consistency often increases its distance from human heuristics, while a random-choice baseline more closely mirrors equitable outcomes when candidates are observationally equivalent. Beyond these average differences, we find that human–AI alignment systematically varies with trial characteristics. As shown in Figure 3, weaker contrasts between applicants, such as small age or experience gaps, correspond to even less machinelike recruiters (around the 10th percentile of the human distribution) to reach 50% agreement with *Skynet*. Sharper contrasts, such as age

gaps above 12 months, raise this threshold toward the 25th percentile of the human distribution of the *machina* trait. We also show a parallel pattern across vacancies: jobs with low education and experience requirements exhibit low location values, meaning that even less machinelike recruiters align with *Skynet*, while skilled vacancies require substantially higher levels of such trait to achieve agreement between men and machine. Taken together, these findings strongly suggest that divergence is not random: the context of the decision systematically shapes both the threshold and the sharpness of human–AI alignment.

Our paper's methods and findings 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 (*20*). Yet advancing that task raises a central question: can such AI agents improve upon human decision-making? 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 such *homo economicus*.

Human hiring referrals are often inconsistent and prone to bias, particularly when individuals from marginalized groups are being evaluated, such as those defined by gender, immigration status, or sexual orientation, as shown in labor market discrimination studies (*5–7, 21, 22*). These limitations can be understood through the lens of bounded rationality, where cognitive constraints lead decision-makers to fail to optimize (*23*). AI systems, with their ability to process large datasets, may move closer to the format of rational "boundlessness" (*24*); however, they still face some constrained capacity concerning processing information, mainly due to finite computing power and incomplete or biased training data (*25, 26*).

Moreover, 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 (*27*). 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 work shows that the differences are substantial and are also affected by context (even within a simplified, yet realistic, candidate screening exercise).

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 behavior (*28, 29*) or when they operate with limited information on productivity-relevant attributes, they risk replicating existing biases under the appearance of objectivity. Moreover, algorithmic decisions may lack legitimacy if perceived as detached from human values or fairness norms (*30, 31*), and human trust in AI varies due to algorithm aversion or appreciation (*32, 33*). Understanding how AI converges with or diverges from human behavior, 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 behavior, therefore, stems from their underlying pretrained models and the specific prompts designed to imbue them with humanlike 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.

Our findings also suggest a core paradox in AI-driven recruitment: the more "AI-like" a re-

cruiter becomes—exhibiting consistency, rule-based reasoning, and alignment with an idealized benchmark like *Skynet*—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 humanlike, they introduce greater variability, possibly undermining the consistency that makes algorithmic systems attractive in the first place.

Our Labor Market Turing Test (LMTT) and cognitive interviewing approach offer a generalizable framework for evaluating AI-human alignment with an eye to the ultimate purpose of ensuring algorithms are consistent, fair, and trustworthy across fields. Importantly, our finding that human recruiters more closely resemble Randomistas—agents who make selections purely at random—underscores an important element regarding fairness. The design of our experiment, according to which candidate pairs are observationally equivalent in terms of productivity and related attributes, implies that a fair decision rule should assign equal probabilities to each candidate's being selected. In this context, the behavior of our Randomista agents, who choose at random with a 50% probability, serves as a normative benchmark for fairness. Paradoxically, what may appear as uninformed randomness represents a form of just behavior under conditions of symmetry and informational equivalence.

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 behaviors closer to this probabilistic fairness—more so than some AI agents optimized for consistency or efficiency— an algorithmic type of decision-making by AI agents that tends to favor certain applicant markers (productive or not) systematically. This framing invites the following reflection: in settings where no observable justification exists for favoring one option over another, true fairness may lie in embracing a degree of randomness.

References and Notes

1. M. Kaufman, Ai might be scanning your resume. here's what job hunters should know (2025). CNN, accessed July 17, 2025.

2. P. Will, D. Krpan, G. Lordan, *Artificial Intelligence Review* **56**, 1071 (2023).

3. M. Hoffman, L. B. Kahn, D. Li, *Quarterly Journal of Economics* **133**, 765 (2018).

4. D. Li, L. Raymond, P. Bergman, Hiring as exploration (2020). Forthcoming, Review of Economic Studies.

5. R. Fabregas, W. Zanoni, *Journal of Public Economics* **247**, 105393 (2025).

6. W. Zanoni, S. Duryea, J. Paredes, *SSRN Electronic Journal* (2024). Accessed: 2025-03-30.

7. W. Zanoni, H. Hernández, O. Zambrano, G. Quiroz, *Labour Economics* **88**, 102531 (2024).

8. G. B. Willis, *Cognitive Interviewing: A Tool for Improving Questionnaire Design* (SAGE Publications, Inc., Thousand Oaks, CA, 2005).

9. X. Liu, *et al.*, *The Lancet Digital Health* **1**, e271 (2019).

10. Vanderbilt University Generative AI Initiative, Prompt patterns, https://www.vanderbilt.edu/generative-ai/prompt-patterns (2025). Accessed: 2025-07-15.

11. A. Turner, The psychology of prompt engineering: Understanding user interaction with ai, https://arsturn.com/blog/the-psychology-of-prompt-engineering-understanding-user-interaction-with-ai (2025). Accessed: 2025-07-15.

12. Microsoft, Prompt engineering concepts in semantic kernel, https://learn.microsoft.com/en-us/semantic-kernel/concepts/prompts (2025). Accessed: 2025-07-15.

13. A. M. Turing, *Mind* **59**, 433 (1950).

14. T. Eriksson, P. Johansson, S. Langenskiöld, *Empirical Economics* **53**, 803 (2017).

15. S. Amershi, *et al.*, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, S. Brewster, G. Fitzpatrick, eds. (ACM, 2019), pp. 1–13.

16. A. Birnbaum, *Statistical Theories of Mental Test Scores*, F. M. Lord, M. R. Novick, eds. (Addison-Wesley, Reading, MA, 1968), pp. 397–479.

17. F. M. Lord, M. R. Novick, *Statistical Theories of Mental Test Scores* (Addison-Wesley, Reading, MA, 1968).

18. S. H. Liu, R.-P. Juster, K. Dams-O'Connor, J. Spicer, *Comprehensive Psychoneuroendocrinology* **3**, 100009 (2020).

19. H. May, *International Journal of Testing* **6**, 1 (2006).

20. D. C. Parkes, M. P. Wellman, *Science* **349**, 267 (2015).

21. R. Bartlett, A. Morse, R. Stanton, N. Wallace, *Journal of Financial Economics* **143**, 30 (2022).

22. J. Dressel, H. Farid, *Science Advances* **4**, eaao5580 (2018).

23. H. A. Simon, *Quarterly Journal of Economics* **69**, 99 (1955).

24. M. Shick, N. Johnson, Y. Fan, *Development and Learning in Organizations: An International Journal* **38**, 1 (2024).

25. M. Jirásek, *Available at SSRN 4703706* (2024).

26. H. Ma, M. Su, *Organizational Dynamics* p. 101100 (2024).

27. J. E. Stiglitz, *The American Economic Review* **65**, 283 (1975).

28. M. Pan, W. Huang, Y. Li, X. Zhou, J. Luo, *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020), pp. 1334–1343.

29. T. Hagendorff, S. Fabi, M. Kosinski, *Nature Computational Science* **3**, 833 (2023).

30. M. K. Lee, S. Baykal (2017), pp. 1035–1048.

31. J. Danaher, *Philosophy and Technology* **29**, 245 (2016).

32. B. J. Dietvorst, J. P. Simmons, C. Massey, *Journal of Experimental Psychology: General* **144**, 114 (2015).

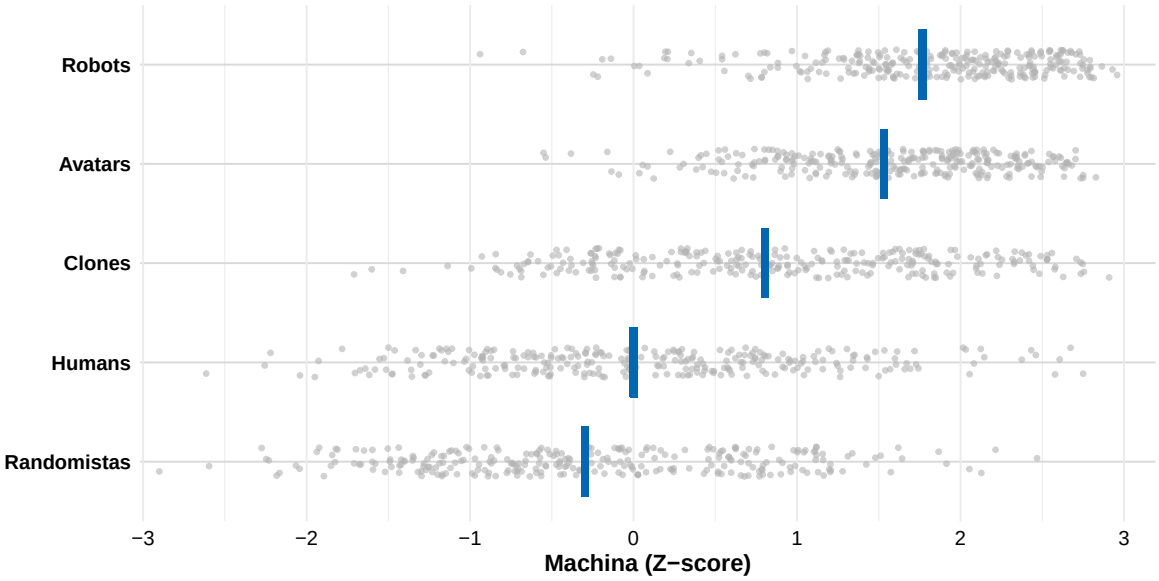
33. J. M. Logg, J. A. Minson, D. A. Moore, *Organizational Behavior and Human Decision Processes* **151**, 90 (2019).

34. A. Shaghaghi, R. S. Bhopal, A. Sheikh, *Health Promotion Perspectives* **1**, 86 (2011).

35. E. F. Wonderlic, *Wonderlic Personnel Test Manual* (Wonderlic Inc., 1999).

36. O. P. John, E. M. Donahue, R. L. Kentle (1991).

37. M. Rosenberg, *Society and the Adolescent Self-Image* (Princeton University Press, 1965).



	Mean	SD	Mean difference test p-value	Variance equality test p-value	Kolmogorov–Smirnov test p-value
Robots	1.77	0.76	0.000	0.000	0.000
Avatars	1.53	0.72	0.000	0.000	0.000
Clones	0.80	0.99	0.000	0.978	0.000
Humans	0.00	1.00	REF	REF	REF
Randomistas	-0.30	0.98	0.000	0.950	0.005

Figure 1: ***Machina*** **by type of recruiter— density functions, descriptive statistics, and basic contrasts.**

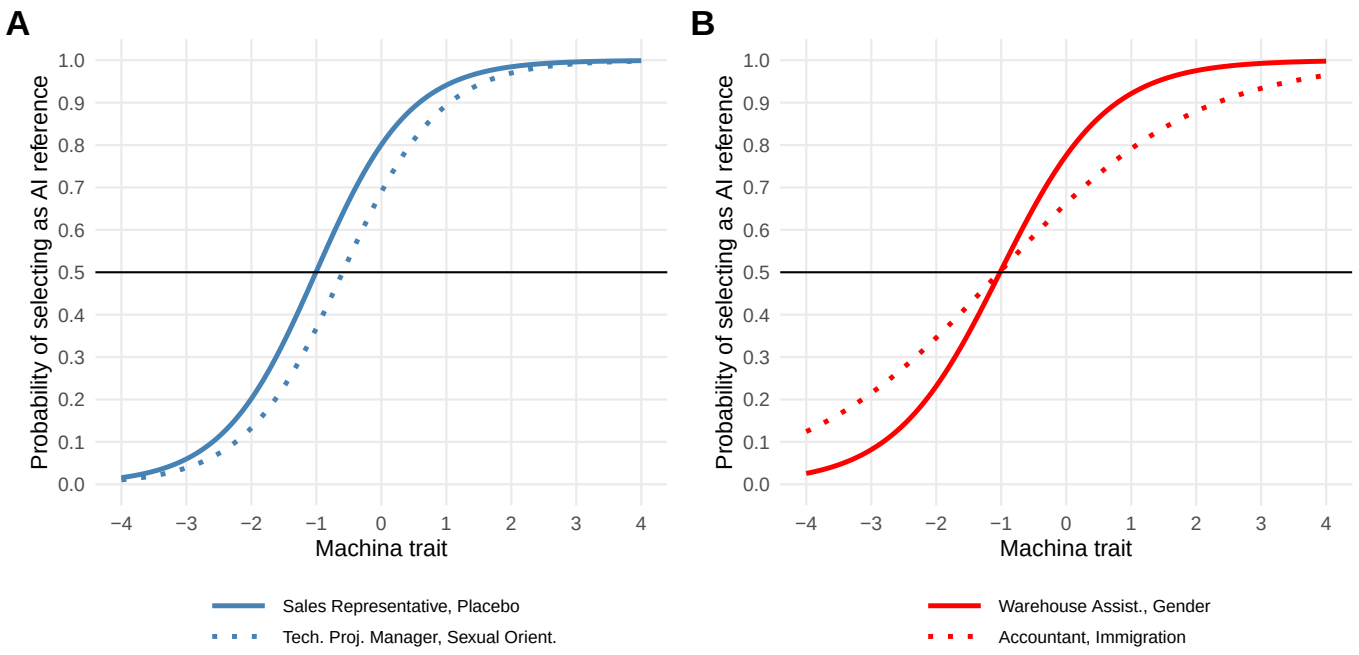


Figure 2: **Characterizing selected cases of screening trials—Trial/item characteristic curves.**

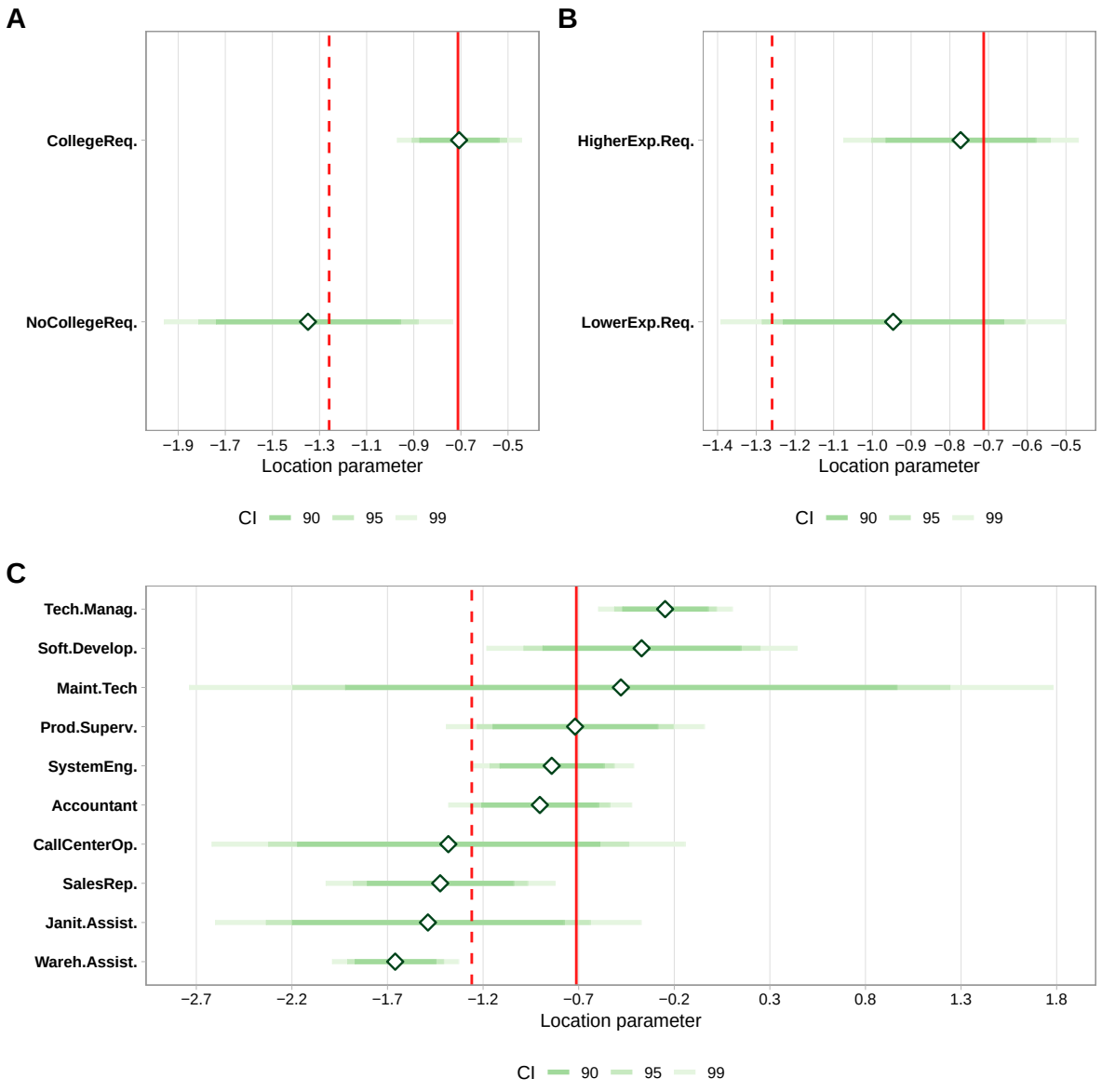


Figure 3: **Location parameter estimates by vacancies' education and experience requirements.** *Notes.* Red vertical solid and dashed lines in Panels A, B, and C mark the 25th and 10th percentiles of the machina latent score among humans, respectively: $P_{25} = -0.712$, $P_{10} = -1.259$.

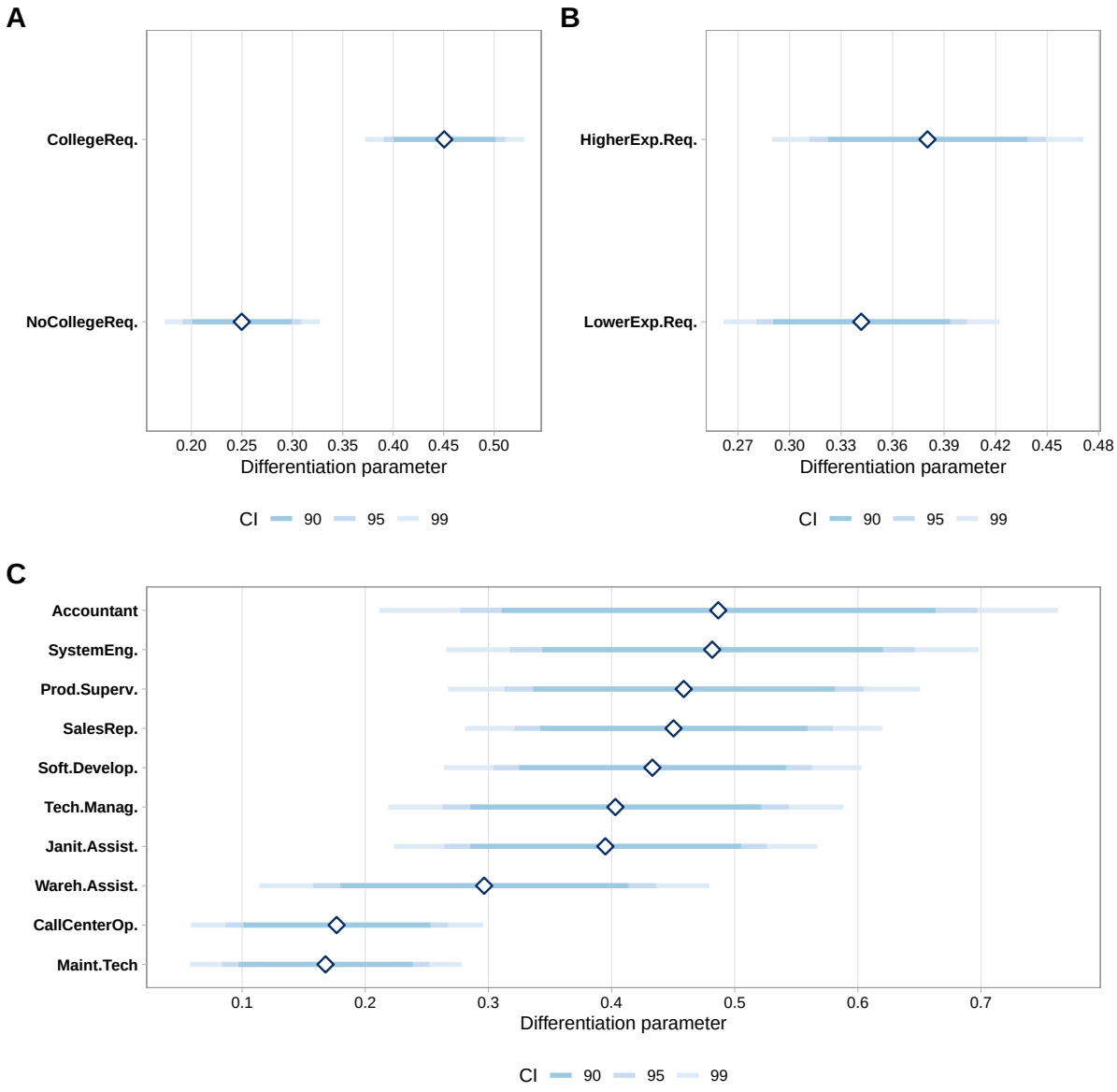


Figure 4: **Differentiation parameter estimates by vacancies' education and experience requirements**

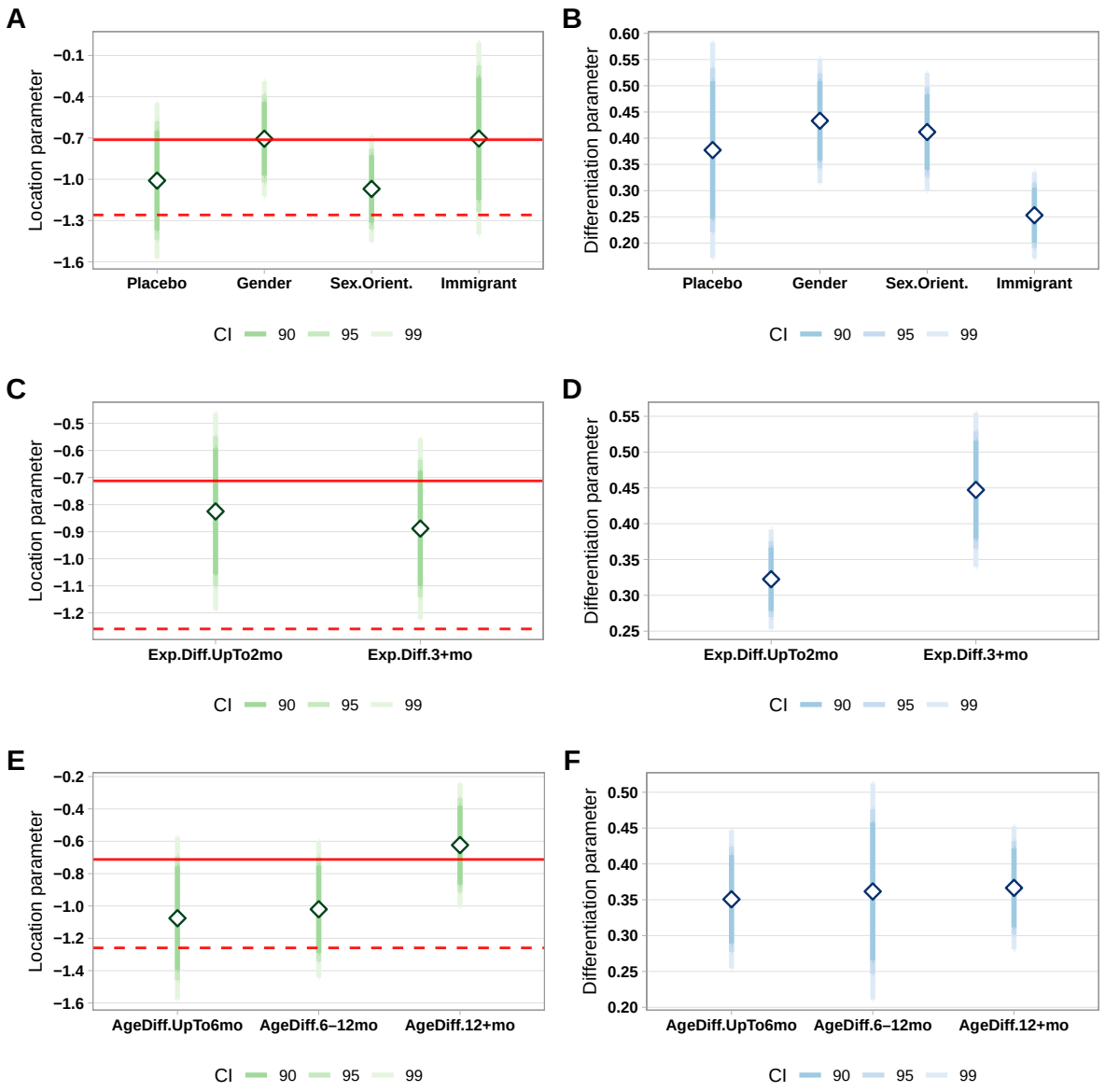


Figure 5: **Parameter estimates by identity themes and by age and experience gaps.** *Notes.* Red horizontal solid and dashed lines in Panels A, C, and E mark the 25th and 10th percentiles of the machina latent score among humans respectively: $P_{25} = -0.712$, $P_{10} = -1.259$.

SM—Supplementary Materials

for *"Documenting Differences Between Humans and AI in High-Stakes Decisions: A Labor Market Turing Test"*

Materials and Methods

SM1. Recruiting human recruiters

We hired experienced human resource recruiters (HRRs) from Quito's labor market. 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 2020 via the operation of a reputable nongovernmental organization with expertise in job search and training programs in Ecuador (*Grupo FARO*). 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 (*34*). 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 Figure SM1. Using different recruitment strategies, the study aimed to capture a diverse range of perspectives and ensure the representation of various segments of the labor market. Recruiters were remunerated competitively (according to the rates of the local market) for their time and expertise.

After manifesting interest, these recruiters were approached via email message (Figure SM2). In this communication recruiters were informed of the two-step process involved in the execution of their freelance task. First, they were asked to respond to a comprehensive questionnaire used to create a detailed profile of their background, including demographic characteristics, work experience, cognitive ability (Wonderlic test), personality traits (Big Five Inventory), and self-esteem (Rosenberg scale). Second, they were asked to take a test of their basic knowledge of the Ecuadorian labor market. The Wonderlic test is a timed 12-minute assessment widely used to evaluate problem-solving skills and general intelligence (*35*). The Big

Five Inventory (BFI) assesses personality based on five key traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism (*36*), while the Rosenberg Self-Esteem Scale is a commonly used 10-item tool that captures individuals' overall perception of self-worth (*37*). We reached 277 recruiters who completed all steps in the experiment.

Ultimately, direct outreach on LinkedIn accounted for 38% of the sample, while peer referral through RDS provided the remaining 62%. Compared those data against national-level labor 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 labor market (*6*).

SM2. Experimental Vacancies

The vacancies in the experiment spanned various sectors with differing required specialization 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 Center Operator. Details of each are presented in Table SM1.

SM3. Experimental 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 randomization of identity markers, a total of four thematic categories were used, with some

represented through multiple variations, resulting in 10 unique paired-CV versions: (1) Nonimmigrant straight woman vs. nonimmigrant straight woman, (2) Nonimmigrant straight woman vs. nonimmigrant straight man (3 versions), (3) Nonimmigrant gay man vs. nonimmigrant straight man (2 versions), (4) Nonimmigrant lesbian vs. nonimmigrant straight woman, (5) Immigrant straight man vs. nonimmigrant straight man, and (6) Immigrant straight woman vs. nonimmigrant straight woman (2 versions).

SM4. AI recruiter teams

To organize team of AI recruiters we used ChatGPT version 4o-mini with an IP address from Ecuador.

AI 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 specialized roles, or follow a random order.

AI 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 labor 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 contextualize 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.

AI Clones: An additional team of 277 Clones was created, each designed to impersonate a specific human recruiter (as an Avatar) but also 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. This design ensured that Clones replicated both the content and the path of decision-making observed in the human data, enabling a highly granular comparison of selection behavior under identical conditions of exposure and sequence.

Randomistas: Our final group of 277 AI recruiters established a purely stochastic baseline. Each Randomista selected a candidate at random in every trial to which they were exposed.

SM5. Central Planner AI—Skynet

Skynet is a unique AI recruiter designed as a baseline to benchmark other recruiting teams' decisions. 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 optimized for fairness. Rather, it reflects how a general-purpose AI system behaves when making referral decisions based solely on prompts that prioritize candidate suitability without any additional targeted constraints or training. Although identity markers were not emphasized in its instructions, *Skynet* may still attend to them implicitly if its underlying training data encoded patterns linking identity to job performance or selection.

SM6. Item Response Theory formulation

Schematically, and without loss of generality, we characterize our model by letting $(c_1 > c_2)_j^S$ denote *Skynet*'s choosing candidate c_1 over c_2 in trial j. Then, for recruiter i, the probability of matching *Skynet* preferred pick is modeled as

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

where μ_i represents the recruiter's latent trait of interest (*machina*). 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 addition to the heterogeneity across individuals in their propensity to agree with the reference machine, 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 we refer to as the "location" parameter (β_j), which serves to locate a given trial along our *machina* trait continuum, indicating the trait level at which recruiters are equally likely to choose or not to choose *Skynet*'s preferred candidate. In other words, the location parameter tells us how much of *machina* a re-

cruiter needs to have a 50/50 shot at choosing like *Skynet*. High values of this parameter indicate that only individuals with very high levels of *machina* would ultimately chose like *Skynet* in that trial. Therefore, our modeling exercise can reveal screening trials that span a broad spectrum of disagreement levels, some associated with high *machina*, others with moderate *machina*, and some with low *machina*.

The second parameter characterizing every trial is referred to as a "differentiation" parameter (α_j), which reflects how effectively the item distinguishes between individuals at different points on the *machina* continuum around the level that is defined by the "location" parameter above. Relatively higher differentiation values indicate that a trial is very sensitive to differences in *machina* trait, and therefore better distinguishes between the machinelike and the nonmachinelike recruiters.

SM7. Regression Specifications for Figures 3, SM8, and SM9

In this section, we describe the models that we employed to estimate the regression coefficients for the location and differentiation parameters presented in Figures 3, SM8, and SM9.

Rescaling the *machina* scores. The raw *machina* scores are not automatically comparable across humans and AI. To make the estimates interpretable in relation to human recruiters, we rescaled the scores so that they are expressed in the same units as the human distribution. Specifically, we take each recruiter's latent *machina* score μ_i, subtract from it the mean of the human distribution θ_{human}, and then divide the result by the human standard deviation σ_{human}. The rescaled score therefore, indicates how far a recruiter lies above or below the human average in standardized units. A value of zero corresponds to the mean human recruiter, positive values indicate recruiters who are more machine-like than the average human (measured in human-standard deviations), and negative values correspond to recruiters who are less machine-like

than the average. Formally, the rescaling is given by:

$$Zmachina_i = \frac{\mu_i - \theta_{human}}{\sigma_{human}},$$

Rescaling the *location* and *differentiation* parameters. Because the latent *machina* score was standardized relative to the human distribution, we also re-scaled the trial-level parameters for consistency. This adjustment ensures that both the *location* and *differentiation* parameters are expressed in the same standardized metric, making them directly comparable to the rescaled *machina* score. For each trial j, the parameters were redefined as:

$$Loc_j = \frac{\beta_j - \theta_{human}}{\sigma_{human}}, \quad Diff_j = \alpha_j \times \sigma_{human}$$

where β_j is the original (unscaled) *location* parameter and α_j the original (unscaled) *differentiation* parameter. The transformed parameters (Loc_j and $Diff_j$) are then used as dependent variables in the regressions. Each regression is estimated over 100 unique trials, with each trial corresponding to a pair of candidate/job vacancy CVs.

Figure 3. Figure 3 examines how identity themes (Panels A–B) and applicant characteristics—work experience gaps (Panels C–D) and age gaps (Panels E–F)—explain variation in the rescaled *location* (Loc_j) and *differentiation* ($Diff_j$) parameters. Each diamond shows a coefficient estimate from an OLS regression, with either Loc_j or $Diff_j$ as the dependent variable. The regressions are estimated over 100 trials, where each trial corresponds to a unique pair of candidate CVs evaluated for a given job vacancy. This design fixes the unit of analysis at the trial level, ensuring that variation in the coefficients reflects systematic differences in applicant characteristics rather than repeated recruiter choices. We estimate a family of OLS regressions without intercepts of the form:

$$y_{pj} \;=\; \mathbf{x}_{pj}^{\top}\beta_p \;+\; \varepsilon_{pj},$$

where $y_{pj} \in \{Loc_j,\ Diff_j\}$ is the dependent variable for trial j in panel p (here panel refers to panels A-F in Figure 3); $\mathbf{x}_{pj}$ is a vector of trial-level indicators (dummies) in an OLS model excluding the constant; and $N = 100$ denotes the number of trials (each trial corresponds to a unique pair of candidate CVs for a given vacancy). The coefficients β_p are interpreted as the mean value of Loc_j or $Diff_j$ associated with the indicators that denote each category, expressed in human-standardized units.

The content of $\mathbf{x}_{pj}$ varies by panel:

- **Panels A–B (Identity themes).** $\mathbf{x}_{pj} = \mathbf{d}_j^{Identity} = [\,Gender_j,\ \textbf{\textit{Placebo}}_j, Sexorient_j,\ Immigrant_j\,]^\top$.
- **Panels C–D (Experience gaps).** $\mathbf{x}_{pj} = \mathbf{d}_j^{ExpGap} = [\,ExpDiff3mo_j, ExpDiffUpTo2mo_j\,]^\top$. Where $ExpDiff3mo_j$ is an indicator for whether the difference in experience between the two candidates is 3 or more months, and $ExpDiffUpTo2mo_j$ being an indicator for whether the difference in experience between the two candidates is up to two months.
- **Panels E–F (Age gaps).** $\mathbf{x}_{pj} = \mathbf{d}_j^{AgeGap} = [\,AgeDiffUpTo6mo_j,\ AgeDiff6To12_j,\ AgeDiff12mo_j\,]^\top$ Where $AgeDiffUpTo6mo_j$, $AgeDiff6To12_j$, and $AgeDiff12mo_j$ are indicators for whether the difference in ages between the two candidates is up to six, between 6 and 12, or more than 12 months, respectively.

Accordingly, Panels A, C, and E report regressions with $y_{pj} = Loc_j$, while Panels B, D, and F report regressions with $y_{pj} = Diff_j$.

Figures SM8 and SM9. Figures SM8 and SM9 examine how vacancy characteristics—education requirements (Panel A), experience requirements (Panel B), and job vacancy type (Panel C)—explain variation in the rescaled parameters. Figure SM8 reports results for the *location* parameter (Loc_j), while Figure SM9 reports results for the *differentiation* parameter ($Diff_j$). Each diamond shows a coefficient estimate from an OLS regression with either Loc_j or $Diff_j$ as the

dependent variable. The regressions are estimated over 100 trials, where each trial corresponds to a unique pair of candidate CVs evaluated for a given job vacancy. Defining the trial as the unit of analysis ensures that the estimated coefficients capture systematic variation across vacancies rather than recruiter-specific effects. We estimate another family of OLS regressions of the form:

$$y_{pj} \;=\; \mathbf{x}_{pj}^{\top}\beta_p \;+\; \varepsilon_{pj},$$

where $y_{pj} \in \{Loc_j,\, Diff_j\}$ is the dependent variable for trial j in panel p (here panel refers to panels A–C in Figures SM8 and SM9); $\mathbf{x}_{pj}$ is a vector of trial-level indicators (dummies); and $N = 100$ denotes the number of trials (each trial corresponds to a unique pair of candidate CVs for a given vacancy). The coefficients β_p are interpreted as the mean value Loc_j or $Diff_j$ associated with the indicators that denote each category, expressed in human-standardized units.

The content of $\mathbf{x}_{pj}$ varies by panel:

- **Panel A (Education requirements).** $\mathbf{x}_{pj} = \mathbf{d}_j^{EduReq} = [\,CollegeReq_j,\, NoCollegeReq_j\,]^{\top}$.
- **Panel B (Experience requirements).** $\mathbf{x}_{pj} = \mathbf{d}_j^{ExpReq} = [\,LowerExpReq_j,\, HigherExpReq_j\,]^{\top}$.
- **Panel C (Job Vacancy Type).** $\mathbf{x}_{pj} = \mathbf{d}_j^{Vacancy} = [\,TechManager_j,\, SoftwareDev._j,\, \ldots\,]^{\top}$.

SM Figures

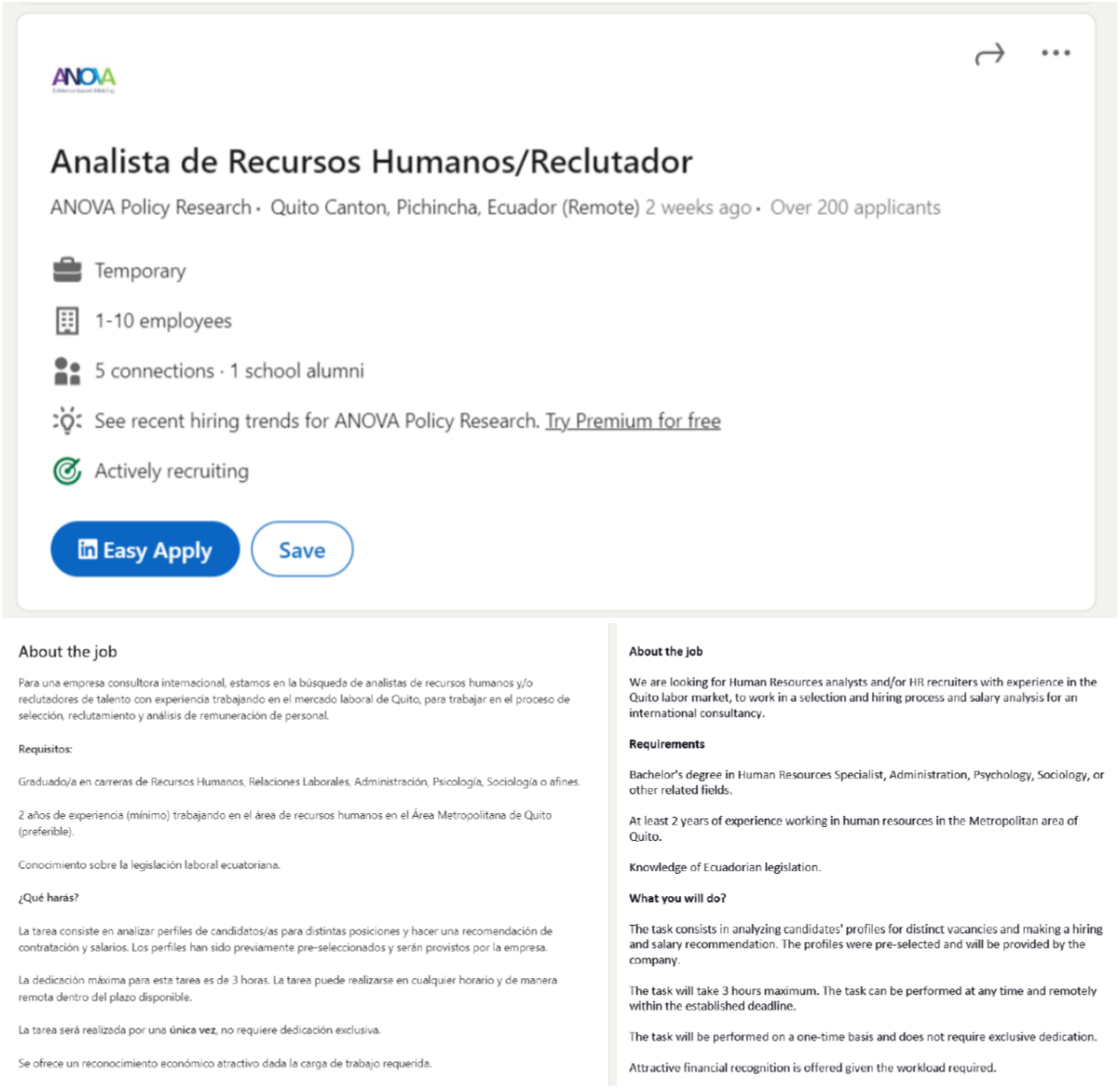


Figure SM1: **Advertisement for human resources recruiters**

Recibe un cordial saludo,

Te saludamos a nombre de ANOVA y FARO, quienes estamos apoyando a una empresa internacional que está pronta a abrir sus actividades en el Ecuador y se encuentra buscando personal. Por esta razón, estamos en la búsqueda de analistas de recursos humanos y/o reclutadores de talento con experiencia en el mercado laboral de Quito para trabajar en el proceso de selección, reclutamiento y análisis de remuneración de personal.

Queremos agradecerte por tu interés para participar como Analista de Recursos Humanos. Nos interesa tu perfil y te hemos preseleccionado para apoyarnos en este proceso de selección.

Este trabajo tiene 2 etapas. La primera consiste en completar tu registro y una segunda que es revisar los perfiles preseleccionados.

Completar todo este proceso tomará menos de 3 horas.

Fase 1: conociendo a nuestros reclutadores

En esta primera fase vamos a pedirte que completes un cuestionario de información general sobre ti y tu experiencia. Te pediremos adicionalmente completar 3 preguntas específicas sobre el mercado laboral ecuatoriano y 2 tests que nos permitirán conocerte mejor.

Completa tu registro durante las 72 horas siguientes a recibir este correo, debido a que luego de ello caducará el link. Este es el enlace de registro:

https://anovarecruiting.net/ecuador/registro/

Fase 2: selección de perfiles

Una vez completado el registro, recibirás un enlace a tu correo que te llevará a la plataforma donde se realiza la selección de los postulantes. En el correo estarán las indicaciones necesarias para empezar con el proceso de selección. Revisa tu correo no deseado, ya que el correo podría llegar ahí.

Te pedimos analizar los perfiles pre-seleccionados de candidatos/as para diversos puestos (10). Te recordamos que esta actividad será desarrollada por una única vez. Esta actividad no requiere dedicación exclusiva y se realizará de manera remota.

El reconocimiento económico por realizar la tarea será de 30 USD que se realizará a través de una transferencia bancaria. Una vez completada la tarea, nuestros aliados de Faro se contactarán contigo para llevar a cabo el pago correspondiente.

Agradecemos tu interés para colaborar con nosotros en este proceso de búsqueda de perfiles adecuados.

Cualquier inquietud o comentario no dudes en escribirnos.

Atentamente,

ANOVA y FARO

We send you our warmest greetings,

We greet you on behalf of ANOVA and FARO, who are supporting an international company that is about to open its activities in Ecuador and is looking for personnel. For this reason, we are looking for human resources analysts and/or talent recruiters with experience in the labor market of Quito to work in the process of selection, recruitment and analysis of remuneration of personnel.

We would like to thank you for your interest in participating as a Human Resources Analyst. We are interested in your profile and we have pre-selected you to support us in this selection process.

This job has 2 stages. The first is to complete your registration and the second is to review the pre-selected profiles.

This whole process will take less than 3 hours to complete.

Phase 1: getting to know our recruiters

In this first phase we will ask you to complete a general information questionnaire about yourself and your experience. We will additionally ask you to complete 3 specific questions about the Ecuadorian labor market and 2 tests that will allow us to get to know you better.

Please complete your registration within 72 hours of receiving this email, because after that the link will expire. This is the registration link:

https://anovarecruiting.net/ecuador/registro/

Phase 2: profile selection

Once the registration is completed, you will receive a link to your email that will take you to the platform where the selection of the applicants is made. The email will contain the necessary instructions to start the selection process. Please check your junk mail, as the email may arrive there.

We ask you to analyze the pre-selected profiles of candidates for various positions (10). We remind you that this is a one-time activity. This activity does not require exclusive dedication and will be performed remotely.

The economic recognition for performing the task will be 30 USD that will be made through a bank transfer. Once the task is completed, our Faro partners will contact you to make the corresponding payment.

We appreciate your interest in collaborating with us in this process of searching for suitable profiles.

If you have any questions or comments, please do not hesitate to contact us.

Yours sincerely,

ANOVA and FARO

Figure SM2: **Communication with human resources recruiters**

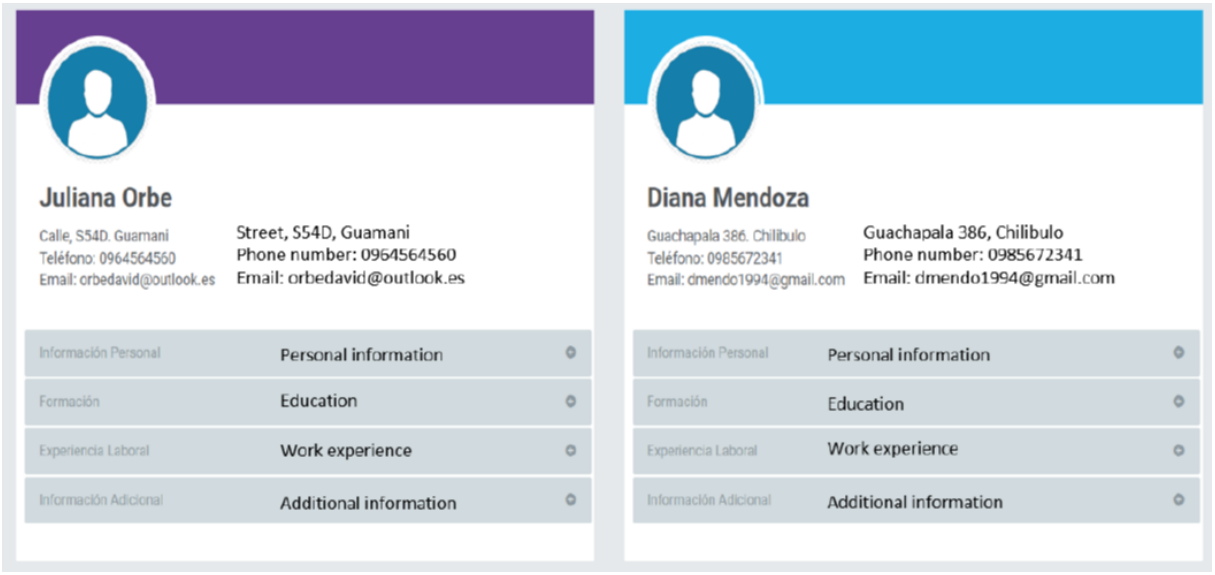


Figure SM3: **Task-management platform interface—Display example**

Figure SM4: **Final candidate selection interface—Display example**

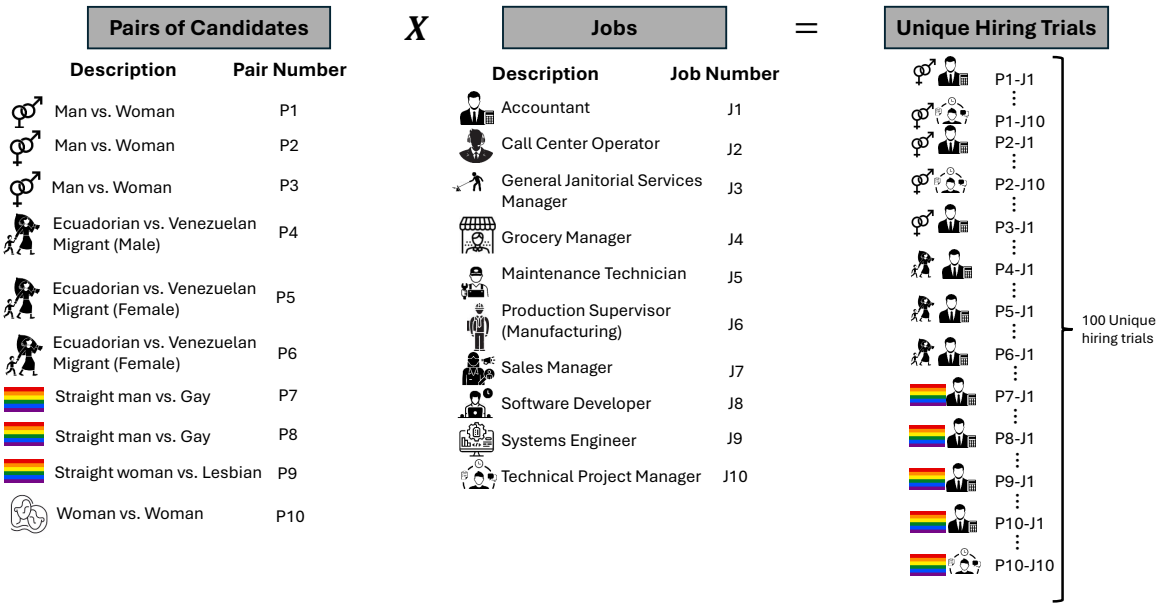
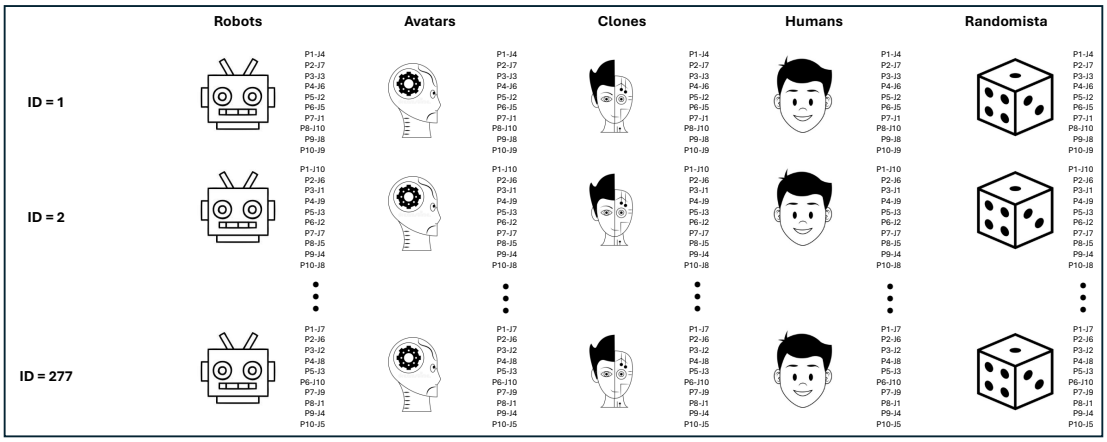


Figure SM5: **Hiring trials**

Recruiters	Hire/Impersonate the Recruiter	Select the Candidate
Robot 		• Read and interpret the HTML program of the web page to evaluate candidates • Select Jobs sequantially 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 *read and interpret* 1. The LinkedIn post or RDS email sent 2. The job instructions sent by email 3. The HTML program of the registration page 4. The recruiters' characteristics Then, they were asked to *impersonate* the recruiters' characteristics	• Read and interpret the HTML program of the web page to evaluate candidates • Select Jobs sequantially 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 *read and interpret* 1. The LinkedIn post or RDS email sent 2. The job instructions sent by email 3. The HTML program of the registration page 4. The recruiters' characteristics Then, they were asked to *impersonate* 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 sequantially 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 SM6: **Human and AI recruiter teams**



Team of Recruiters	Unique Hiring Trial 1	Unique Hiring Trial 2	Unique Hiring Trial 3	Unique Hiring Trial 4	...	Unique Hiring Trial 100
Robots	P1-J1	P1-J2	P1-J3	P1-J4	...	P10-J10
Avatars	P1-J1	P1-J2	P1-J3	P1-J4	...	P10-J10
Clones	P1-J1	P1-J2	P1-J3	P1-J4	...	P10-J10
Humans	P1-J1	P1-J2	P1-J3	P1-J4	...	P10-J10
Randomista	P1-J1	P1-J2	P1-J3	P1-J4	...	P10-J10

Figure SM7: **Unfolding of the experiment: The tasks of humans and AI recruiters**

SM Tables

Table SM1: Full vacancy descriptions used in the screening experiment

Job Opening	General Objective	Key Responsibilities	Technical Skills Required	Educational Requirements	Company	Location
Sales Representative	Experienced commercial professional to support the company's expansion in the Ecuadorian market.	Commercial advising, client prospecting, visits, follow-ups, sales closing, and collections.	CRM, computing, databases, office, sales, customer service.	University degree in business or management. 3 years of experience.	Food Industry	Quito
Janitorial Services Assistant	Responsible for cleaning offices and common areas.	Cleaning of offices, restrooms, storage areas; sanitization.	Cleaning supplies and equipment.	High school. Experience desirable.	Construction	Quito
Warehouse Assistant	Manage procurement, storage, and delivery of materials and goods.	Update materials list, handle quotes, verify deliveries.	Office, inventory handling, email.	Technical degree in warehouse management. 2 years of experience.	Food Industry	Quito
Certified Public Accountant (CPA)	Finance and accounting professional for company operations.	Client service, tax filings, audits, financial statements.	SIAP, Excel, and accounting platforms.	Degree in accounting or related fields. 3 years of experience.	Accounting Firm	Quito
Software Developer	Collaborate with the IT team to develop solutions for internal requirements.	Validate requirements, design solutions, develop software.	Java, .NET, databases, SAP, web development.	Degree in software or systems engineering. 2 years of experience.	Technology	Quito
Systems Engineer	IT professional for backend/frontend control.	AWS implementation, functional programming, data infrastructure.	REST, SCRUM, HTML, NodeJS, microcontroller programming.	Degree in systems or computer science. 4 years of experience.	Technology	Quito
Technical Project Manager	Lead the technical team in nationwide operations.	Indicator control, second-level tech support, process improvement.	CISCO certification, networks, tech support.	Degree in telecommunications. 4 years of experience, 2 in management.	Technology	Quito
Call Center Operator	Handle phone-based client management and campaigns.	Client interaction, campaign handling, follow-up.	Communication, multitasking, sales, portfolio management.	Technical degree in administration. 2 years of experience.	Services	Quito
Production Supervisor (Manufacturing)	Oversee production scheduling and supervision.	Production planning, personnel supervision, process control.	AutoCAD, welding, industrial processes.	Degree in industrial or mechanical engineering. 2 years of experience.	Engineering	Quito
Maintenance Technician	Perform preventive and corrective maintenance.	Electrical/mechanical review, calibration, system improvements.	PLC, programming, electrical control.	Technical or engineering degree in mechanics, electricity, electronics. 3 years preferred.	Engineering	Quito

Table SM2: **Human recruiters—descriptive statistics**

	mean	sd	skewness	p25	p50	p75
Demographics and Education						
Recruited via LinkedIn (=1)	0.36	0.48	0.56	0.00	0.00	1.00
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
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
Work experience						
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
Socio-emotional and Cognitive Scores						
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
Observations	277					

Table SM3: **Trials—descriptive statistics**

	Mean
Education and Experience Requirements	
No College Required	0.40
College Required	0.60
Low Experience Required	0.10
Med-Low Experience Required	0.40
Med-High Experience Required	0.30
High Experience Required	0.20
Theme	
Theme: Placebo	0.10
Theme: Gender	0.30
Theme: Sexual orientation	0.30
Theme: Immigration	0.30
CV's Age & Experience gap	
Diff. Age up to 6 months	0.32
Diff. Age 6-12 months	0.24
Diff. Age 12+ months	0.44
Diff. Exp. up to 2 months	0.66
Diff. Exp. 3+ months	0.34
Observations	100

Table SM4: **Machina trait versus covariates—OLS and Quantile Regressions**

	OLS Regression	Quantile Median	Bottom Third	Top Third
Demographics and Education				
Male (=1)	-0.044	0.029	-0.013	-0.131
	(0.747)	(0.817)	(0.923)	(0.380)
Age (years)	-0.015	-0.025***	-0.029***	-0.008
	(0.175)	(0.010)	(<0.001)	(0.559)
Postgraduate degree (=1)	0.369	0.524*	0.687***	0.437*
	(0.124)	(0.050)	(0.010)	(0.067)
College degree (=1)	0.000	-0.024	-0.015	0.115
	(0.999)	(0.921)	(0.942)	(0.548)
HR-related major (=1)	0.062	0.111	0.014	0.214
	(0.715)	(0.492)	(0.937)	(0.255)
Work experience				
Currently employed (=1)	0.027	0.000	0.064	-0.173
	(0.833)	(0.998)	(0.638)	(0.284)
HR-related current/last job (=1)	-0.193	-0.340	-0.417	-0.258
	(0.392)	(0.115)	(0.129)	(0.379)
HR experience abroad (=1)	-0.003	0.057	-0.037	0.057
	(0.987)	(0.805)	(0.868)	(0.764)
High knowledge of Ecuadorian market (=1)	-0.011	-0.095	-0.110	-0.017
	(0.947)	(0.638)	(0.490)	(0.867)
Socio-emotional and Cognitive Scores				
Neuroticism (z-score)	-0.150*	-0.238***	-0.031	-0.193*
	(0.054)	(0.003)	(0.726)	(0.083)
Extroversion (z-score)	0.029	0.102	0.025	0.094
	(0.744)	(0.266)	(0.779)	(0.315)
Openness (z-score)	-0.056	-0.001	-0.060	0.008
	(0.447)	(0.988)	(0.452)	(0.915)
Agreeableness (z-score)	0.051	0.014	0.008	-0.015
	(0.585)	(0.876)	(0.925)	(0.874)
Conscientiousness (z-score)	-0.038	-0.014	-0.039	-0.126
	(0.642)	(0.855)	(0.652)	(0.165)
Self-esteem (z-score)	0.138**	0.205***	0.136***	0.238**
	(0.028)	(<0.001)	(0.009)	(0.010)
Cognitive test (z-score)	0.008	0.061	0.051	0.044
	(0.902)	(0.394)	(0.385)	(0.556)
Observations	277	277	277	277
R^2	0.087			
Pseudo-R^2		0.095	0.098	0.074

Note: *p*-values in parentheses. All models include fixed effects for sector of professional experience. Observations are weighted by the precision of the latent trait estimated for each recruiter. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table SM5: **Parameters by education and experience requirements—All models**

	Location parameter					Differentiation parameter				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Education and Experience Requirements										
No College Required	REF		REF			REF		REF		
College Required	0.670**		0.697*			0.217***		0.219***		
	(0.048)		(0.063)			(0.000)		(0.000)		
Lower Experience Required	REF		REF	REF	REF	REF		REF	REF	
Higher Experience Required	-0.058		0.180	-0.096	-0.053	-0.042		-0.040	-0.271***	-0.022
	(0.824)		(0.557)	(0.868)	(0.870)	(0.340)		(0.384)	(0.003)	(0.758)
Theme										
Theme: Placebo		REF	REF	REF	REF		REF	REF	REF	REF
Theme: Gender		0.337	0.444*	2.337**	0.176		0.048	0.041	0.134*	-0.056
		(0.249)	(0.062)	(0.024)	(0.418)		(0.580)	(0.606)	(0.094)	(0.660)
Theme: Immigration		0.406	0.557	4.099***	-0.047		-0.147*	-0.126	-0.096	-0.179
		(0.259)	(0.101)	(0.002)	(0.871)		(0.077)	(0.102)	(0.225)	(0.152)
Theme: Sexual orientation		-0.063	0.034	1.632*	-0.125		0.030	0.026	0.113	-0.051
		(0.825)	(0.880)	(0.098)	(0.539)		(0.721)	(0.742)	(0.111)	(0.688)
CV's Age & Experience gap										
Diff. Age up to 6 months		REF	REF	REF	REF		REF	REF	REF	REF
Diff. Age 6-12 months		0.087	0.532**	2.256*	0.573**		-0.100	0.035	-0.119	0.302**
		(0.720)	(0.036)	(0.091)	(0.023)		(0.125)	(0.620)	(0.125)	(0.014)
Diff. Age 12+ months		0.527**	0.707**	2.849**	0.498		-0.073	-0.018	0.064	0.003
		(0.031)	(0.016)	(0.010)	(0.105)		(0.204)	(0.741)	(0.492)	(0.964)
Diff. Exp. up to 2 months		REF	REF	REF	REF		REF	REF	REF	REF
Diff. Exp. 3+ months		0.136	-0.018	0.951	-0.460**		0.115**	0.058	0.236***	-0.047
		(0.534)	(0.943)	(0.155)	(0.030)		(0.021)	(0.252)	(0.000)	(0.470)
Observations	100	100	100	40	60	100	100	100	40	60
R^2	0.061	0.073	0.151	0.349	0.177	0.160	0.149	0.276	0.531	0.163

Note: p-values in parentheses. Equations (4) and (9) are estimated on the subset of trials with no college requirement, while equations (5) and (10) are estimated on the subset of trials with a college requirement. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.