International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Scoring the Score: An Empathy, Trust and Accountability Lens for Auditing AI-Assisted Hiring Decisions

Joshua Fernandes

Indian Institute of Management, Sambalpur, India
Email: *joshua.1234511[at]yahoo.in*
ORCID: 0009-0007-7988-5014

Abstract: ***Hiring is now frequently mediated by algorithms that read résumés, grade recorded interviews, and rank applicants. Such systems are commonly evaluated on a single metric- how accurately they forecast performance. This paper argues that accuracy is a necessary but incomplete test, because it says nothing about how the resulting score is experienced by the applicant, understood by the recruiter, or defended before a regulator. We propose a three-part evaluation lens- Empathy, Trust and Accountability (ETA)- that assesses the decision the machine produces rather than only the model that produced it. Empathy concerns the dignity of the candidate's experience; Trust concerns whether the score is explainable and job-relevant; Accountability concerns whether it is bias-audited, lawful and open to human override under regimes such as the EU AI Act, NYC Local Law 144 and India's DPDP Act 2023. We express the lens as a small structural model with eight propositions, embody it in a free, open-source evaluation module, and illustrate it on a video-interview scenario. As future work, a survey-and-interview study is specified to test the propositions with accuracy held constant. The result is a conceptual framework and open-source artifact- an audit-ready instrument for responsible hiring AI.***

Keywords: AI hiring, algorithmic fairness, explainable AI, AI governance, organisational justice

1. Introduction

A growing number of job-seekers are now first assessed not by a person but by a program. Software shortlists curriculum vitae, awards marks to recorded video answers for traits such as "communication" or "drive", and prepares the polite refusals that no manager ever signs [1], [2]. If a rejected applicant were to ask the recruiter to justify the outcome, the candid reply in most firms would be that the reasoning sits inside a model no one present can read. Efficiency has increasingly been prioritised over explanation at the entrance to the organisation.

Buyers of these systems ask overwhelmingly for one thing: *predictive validity*, the degree to which yesterday's scores matched who later performed. That question matters, and selection research has pursued it for a century. It is, however, only one of the questions worth asking. A score can forecast well and still be produced unfairly, communicated coldly, and left impossible to justify. Accuracy describes how well a model *guesses*; it is mute on whether the guess was *decent*, *legible* or *defensible*. Treating a good accuracy figure as a licence to deploy conflates competence with trustworthiness.

That conflation has become a legal exposure, not merely an ethical worry. The European Union now treats recruitment and selection AI as "high-risk", attaching duties of transparency, oversight, data governance and risk management [3]. New York City compels yearly independent bias audits of automated hiring tools and prior notice to candidates [4]. India's Digital Personal Data Protection Act, 2023, hands applicants enforceable control over their personal data [5]. What each regime demands- an explanation, an audit trail, respectful handling, a human in the loop- are attributes of the *output and the process that yields it*, and are invisible to a validity coefficient.

This paper offers an instrument for exactly those attributes. Building on earlier studies of empathetic machine decision-making [6], [7], we set out an Empathy–Trust–Accountability (ETA) lens that grades the decision itself along three human-centred axes, keeping predictive accuracy fixed as a control. Our contribution is threefold: (i) we separate two acts of scoring - the tool grading the candidate, and a governance layer grading the tool- and argue the second is where responsibility lives; (ii) we knit stakeholder, signalling, acceptance, justice and accountability theory into one measurable lens with falsifiable propositions; and (iii) we ship the lens as runnable open-source software, so a claim of fairness can be executed and checked rather than merely announced.

2. Literature Survey

AI in staffing. Reviews of algorithmic human-resource management document rapid uptake alongside thin governance, warning that opaque scoring imports data-quality and legitimacy risks the field has not resolved [1]. Empirical audits of commercial hiring tools find that fairness and validity claims are frequently unverifiable from the outside [2].

Explaining machine judgments. The explainable-AI literature distinguishes post-hoc rationales from intrinsically interpretable models and cautions that a plausible-looking explanation is not the same as a faithful one [8], [9]; some authors argue that in consequential settings interpretable models should be preferred outright [10]. For a recruiter who must justify a rejection, an explanation is not decoration but the precondition of acting responsibly.

Whether people use the tool. Technology-acceptance research shows adoption turns on perceived usefulness and

ease, later broadened into unified acceptance models [11], [12], [13]. Two opposing failure modes recur: users defer blindly to a confident number, or abandon a useful model after a single visible mistake- automation bias and algorithm aversion respectively [14]. Both are calibration failures that only a reason-bearing score can correct. Studies of automated and augmented decision-making confirm that removing the human from the loop reshapes how the decision is received by workers and observers alike [15], and applicants report lower fairness for AI-run interviews than for human ones, driven by reduced voice and interpersonal warmth rather than by the verdict [16].

Fairness and justice. The organisational-justice tradition explains applicant reactions through procedural, distributive and interactional fairness [17], [18], [19], while stakeholder [20] and signalling theory [21], [22] frame how the *treatment* of a candidate signals the organisation's character to a watching labour market. Accountability theory adds the demand that a decision be answerable to an authority [23], and the ethics-of-algorithms literature maps the transparency, bias and responsibility concerns that follow [24]. Critically, scholars warn against "solving" fairness purely inside the model, since a metric optimised in training abstracts away the social context in which the decision actually lands [25], [26].

Law as a design brief. Article 22 of the GDPR constrains solely-automated significant decisions [27], though the practical scope of any "right to explanation" is contested [28]; the US Uniform Guidelines' four-fifths rule anchors adverse-impact testing [29]; and the EU, NYC and Indian regimes cited above convert these principles into concrete duties. Across sources the obligations converge on respect, explanation and auditable accountability — precisely the axes no standard validation report measures.

3. Problem Definition

The gap this paper addresses can be stated plainly. Existing evaluation of AI-assisted hiring measures a single property of the *model*- predictive validity- whereas the emerging legal and ethical expectations attach to three properties of the *decision*: whether it respects the candidate, whether it can be explained to and defended by the recruiter, and whether it can be audited and overseen under law. No widely adopted, operational instrument scores those three properties on a per-decision basis. We therefore ask:

- **RQ1.** Can the empathy, explainability and accountability of an AI hiring decision be scored reliably and separately from its predictive accuracy?
- **RQ2.** Do those scores predict two outcomes that determine responsible adoption- recruiter acceptance of the tool and candidate perceptions of fairness- beyond what accuracy alone explains?
- **RQ3.** Can the instrument be delivered as auditable software that an organisation, a regulator or an auditor can actually run?

4. Methodology / Approach

4.1 Two acts of scoring

We separate two things that the word "score" conflates. The *first* act is existing practice: an AI grades a candidate and emits a fit value or rank. The *second* act- our object of study- grades that grading, asking how the machine reached and conveyed its judgment. The ETA lens governs the second act. Throughout, predictive validity is a held-constant control, so that any movement in the outcomes can be attributed to the human-centred axes rather than to accuracy. Table 1 defines the three axes and their theoretical anchors, and Figure 1 renders the model with each proposition placed on a path.

4.2 The three axes and propositions

Empathy (candidate experience). Rooted in stakeholder [20] and signalling theory [21], [22] and in the justice account of applicant reactions [17], empathy asks whether the interaction and any feedback treat the applicant as a person with context- an accent, a caregiving gap, a disability- rather than a row in a table.

P1. Empathy is positively related to candidate-perceived fairness, controlling for accuracy and outcome favourability.

P2. Empathy is positively related to recruiter acceptance, via recruiters' anticipation of candidate reactions.

Trust (score explainability). Grounded in explainable-AI [8], [9], [10] and acceptance theory [11], [12], [13], trust asks whether the score is transparent, tied to job-relevant factors and calibrated for the person who must defend it.

P3. Explainability is positively related to recruiter acceptance, via perceived usefulness and trust.

P4. Explainability *moderates* the accuracy–acceptance link: accuracy earns acceptance only when the score can be explained.

Accountability (fairness by audit). Anchored in justice [18], [19] and accountability theory [23] and in the audit logic of Local Law 144 and the four-fifths rule [4], [29], accountability asks whether the score is traceable, bias-audited, lawful and open to override.

P5. Accountability is positively related to candidate-perceived fairness, chiefly its procedural and interactional facets.

P6. Accountability is positively related to recruiter acceptance in regulated settings, via perceived risk reduction.

P7. Accountability *moderates* the accuracy–acceptance link: accuracy earns acceptance only when the decision can be audited.

P8. Accuracy alone is insufficient; acceptance and perceived fairness are jointly produced by empathy, explainability and accountability, so a high-accuracy tool that scores poorly on ETA will not sustain lawful adoption.

4.3 A compact formal model

For output i the instrument returns empathy E_i, trust T_i and accountability A_i on a 1–5 scale, plus the fixed accuracy control PV_i. A composite index summarises the three human axes,

$$\mathrm{ETA}_i = w_E E_i + w_T T_i + w_A A_i, \quad w_E + w_T + w_A = 1, \quad (1)$$

with equal weights by default and regime-specific weights (for instance, favouring accountability where the law is strict) as an option. The two outcomes — acceptance (ACC) and perceived fairness ($FAIR$) — are modelled as

$$ACC = \beta_0 + \beta_1 PV + \beta_2 E + \beta_3 T + \beta_4 A + \beta_5 (PV \times T) + \beta_6 (PV \times A) + \varepsilon, \quad (2)$$

$$FAIR = \gamma_0 + \gamma_1 E + \gamma_2 A + \gamma_3 PV + \zeta. \quad (3)$$

Each proposition becomes a sign test on a coefficient (Table 2), so the lens is refutable rather than merely descriptive. Accountability is tied to statute through the adverse-impact ratio,

$$AIR = \frac{SR_p}{SR_r}, \quad \text{flag adverse impact if } AIR < 0.80, \quad (4)$$

where SR_p and SR_r are selection rates for a protected and a reference group; the accountability axis rewards outputs whose reasoning is auditable against (4) and penalises reliance on proxies for protected traits.

4.4 The software artifact

Following design-science method [30], [31], the lens is realised as a free and open-source evaluation module that extends an existing empathy-scoring pipeline through two hooks- one that shapes the evaluation prompt, one that appends scores. It adds two governed metrics, *score explainability* and *fairness/accountability*, each rated 1–5 by a language model against an explicit rubric tied to a selectable regime (EU AI Act, LL144, DPDP or GDPR). Base empathy supplies the candidate-experience axis and a decision-accuracy metric supplies the fixed control. Evaluation is gated so the governed metrics fire only on hiring scenarios, and the module ships three: video-interview grading, CV screening and candidate feedback. Every run exports per-decision scores for later statistical analysis.

4.5 Planned empirical validation

We specify, but do not execute here, a study to test the propositions with accuracy held constant. HR professionals, recruiters and hiring managers form the primary sample (initially N ≈ 20–30 for instrument refinement, scaling to N ≈ 150–300), with an optional candidate sub-sample for the fairness side. A vignette experiment holds the candidate and the accuracy figure fixed while varying explanation (present/absent) and a bias-audit statement (present/absent). Outcomes are recruiter acceptance (acceptance-model items) and perceived fairness (procedural, distributive, interactional). Following established survey practice, each construct will be measured with a structured questionnaire whose content validity and reliability are established before use- the same instrument-development discipline demonstrated in prior IJSR fieldwork on human-centred outcomes [32]. Analysis uses partial least squares structural equation modelling for the propositions and thematic analysis of 8–12 interviews for mechanism; hypotheses will be pre-registered. Because the governed metrics are scored by a language model, agreement between the model's scores and independent ratings from human experts will be quantified on a held-out sample using the intraclass correlation coefficient (ICC) for the 1–5 ratings and Krippendorff's α for ordinal agreement, with a pre-specified threshold (e.g., $\alpha \geq 0.70$) required before the automated scores are treated as reliable. Only synthetic vignettes and perception data are collected, under institutional ethics approval and informed consent.

5. Results and Discussion

Because this paper contributes a framework and an artifact rather than field data, its "results" are the instrument and a demonstration of its behaviour. Applied to a video-interview vignette- a capable candidate with a regional accent and an employment gap, both non-job-relevant and common bias proxies- the module distinguishes two AI outputs that carry the *same* accuracy. An opaque "score-and-rank" output scores low on trust and accountability; a reason-bearing output that names and discounts the irrelevant factors and cites the applicable oversight duty scores high (Figure 2). The module also withholds the governed metrics from non-hiring scenarios, evidence that the axes are separable and the rubric discriminates. These runs establish feasibility, not human-subject validity, which the planned study will supply.

Three implications follow. *Theoretically*, responsibility is relocated from the model to the decision artefact, reconciling the validity tradition with the justice and explanation demands of stakeholders and law, and the moderation propositions make explicit a claim usually left implicit- that accuracy converts to acceptance only through explanation and audit. *Practically*, organisations facing the EU AI Act, LL144 and the DPDP Act gain an audit-ready procedure that interrogates a tool rather than trusting a vendor's assurance; because the code is open, the rubric is inspectable and contestable. *For the field*, per-decision ETA scores give auditors and regulators a tractable evidence trail aligned with statutory audit logic, and reorient vendor competition from accuracy claims toward demonstrable explainability and fairness.

The main limitations are that the propositions await empirical test, that language-model scoring introduces evaluator dependence requiring inter-rater checks against human experts, and that the instrument is tuned to 2023–2024 regimes and to hiring specifically.

6. Conclusion

Well-governed AI could widen access to work and strip prejudice out of screening rather than bake it in- but only if we ask of a machine what we ask of any interviewer: treat people decently, show your reasoning, and answer for your errors. The ETA lens makes those demands measurable and the open-source module makes them executable. A rejection that no human can explain is not a modern hiring process; it represents an inadequately governed one. Grading the score

itself- on empathy, trust and accountability- is how a firm can put AI at the front door of employment and still stand behind the outcome. As a conceptual framework and working artifact, the ETA lens provides an operational foundation for future empirical validation and practical deployment; large-scale validation remains future work.

7. Future Scope

Future work will run the specified survey and interview study and report the structural model; extend the lens to other selection stages such as sourcing and offer negotiation; test cross-domain transfer against a sibling instrument built for clinical decision support; benchmark scoring stability across different language-model providers; and track the instrument's alignment as the EU AI Act's implementing rules and comparable regulations mature.

References

[1] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," *California Management Review*, vol. 61, no. 4, pp. 15–42, 2019.
[2] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proc. 2020 Conf. on Fairness, Accountability, and Transparency*, New York: ACM, 2020, pp. 469–481.
[3] European Parliament and Council of the European Union, *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, Official Journal of the European Union, 2024.
[4] New York City Council, *Local Law 144 of 2021: Automated Employment Decision Tools*, New York City Administrative Code, 2021.
[5] Government of India, *The Digital Personal Data Protection Act, 2023*, New Delhi: The Gazette of India, 2023.
[6] J. Fernandes, "Evaluating empathetic decision-making in AI: A comparative study of open-source models in high-stakes scenarios," *International Journal for Multidisciplinary Research*, vol. 7, no. 6, 2025.
[7] J. Fernandes, "Operationalising empathy in AI systems for high-stakes decision-making," *International Journal of Innovative Science and Research Technology*, 2026.
[8] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[9] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
[11] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.
[12] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003.
[13] V. Venkatesh, J. Y. L. Thong, and X. Xu, "Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology," *MIS Quarterly*, vol. 36, no. 1, pp. 157–178, 2012.
[14] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology: General*, vol. 144, no. 1, pp. 114–126, 2015.
[15] M. Langer and R. N. Landers, "The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers," *Computers in Human Behavior*, vol. 123, art. 106878, 2021.
[16] Y. Acikgoz, K. H. Davison, M. Compagnone, and M. Laske, "Justice perceptions of artificial intelligence in selection," *International Journal of Selection and Assessment*, vol. 28, no. 4, pp. 399–416, 2020.
[17] S. W. Gilliland, "The perceived fairness of selection systems: An organizational justice perspective," *Academy of Management Review*, vol. 18, no. 4, pp. 694–734, 1993.
[18] J. A. Colquitt, "On the dimensionality of organizational justice: A construct validation of a measure," *Journal of Applied Psychology*, vol. 86, no. 3, pp. 386–400, 2001.
[19] J. Greenberg, "Organizational justice: Yesterday, today, and tomorrow," *Journal of Management*, vol. 16, no. 2, pp. 399–432, 1990.
[20] R. E. Freeman, *Strategic Management: A Stakeholder Approach*. Boston: Pitman, 1984.
[21] M. Spence, "Job market signaling," *Quarterly Journal of Economics*, vol. 87, no. 3, pp. 355–374, 1973.
[22] B. L. Connelly, S. T. Certo, R. D. Ireland, and C. R. Reutzel, "Signaling theory: A review and assessment," *Journal of Management*, vol. 37, no. 1, pp. 39–67, 2011.
[23] M. Bovens, "Analysing and assessing accountability: A conceptual framework," *European Law Journal*, vol. 13, no. 4, pp. 447–468, 2007.
[24] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, pp. 1–21, 2016.
[25] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proc. Conf. on Fairness, Accountability, and Transparency*, New York: ACM, 2019, pp. 59–68.
[26] J. Sánchez-Monedero, L. Dencik, and L. Edwards, "What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems," in *Proc. 2020 Conf. on Fairness, Accountability, and Transparency*, New York: ACM, 2020, pp. 458–468.
[27] European Parliament and Council of the European Union, *Regulation (EU) 2016/679 (General Data*

Protection Regulation), Official Journal of the European Union, L119, pp. 1–88, 2016.
[28] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017.
[29] Equal Employment Opportunity Commission, *Uniform Guidelines on Employee Selection Procedures*, 29 C.F.R. Part 1607, Washington, DC, 1978.
[30] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.
[31] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.
[32] M. A. Pereira, "Effectiveness of child-to-child approach in imparting knowledge regarding road traffic safety among scholars in selected schools of Goa," *International Journal of Science and Research (IJSR)*, vol. 10, no. 6, June 2021, Paper ID: SR21604124626, ISSN: 2319-7064.

Author Profile

Joshua Fernandes is a doctoral researcher in human-centred AI governance at the Indian Institute of Management, Sambalpur. He develops audit-ready methods for evaluating high-stakes AI decisions on empathy, trust and accountability, released as open-source software, and has published earlier peer-reviewed work on empathetic decision-making in open-source AI models.

Tables

Table 1: The three ETA axes and their anchors

Axis	Question the audit asks	Theory	Module metric (1–5)
Empathy	Does the interaction respect the candidate's dignity and context?	Stakeholder; signalling; justice reactions	Empathy
Trust	Is the score transparent, job-relevant and calibrated for the recruiter?	Explainable AI; acceptance models	Score explainability
Accountability	Is the score traceable, bias-audited, lawful and overseeable?	Justice; accountability; LL144 / four-fifths	Fairness / accountability
Accuracy (control)	Does the model predict performance well?	Selection psychometrics	Decision accuracy (fixed)

Table 2: Propositions as coefficient sign-tests on Equations (2)–(3)

Proposition	Path	Predicted sign	Outcome
P1	Empathy → Fairness	$\gamma_1 > 0$	Perceived fairness
P2	Empathy → Acceptance	$\beta_2 > 0$	Acceptance
P3	Trust → Acceptance	$\beta_3 > 0$	Acceptance
P4	Accuracy × Trust → Acceptance	$\beta_5 > 0$	Acceptance
P5	Accountability → Fairness	$\gamma_2 > 0$	Perceived fairness
P6	Accountability → Acceptance	$\beta_4 > 0$	Acceptance
P7	Accuracy × Accountability → Acceptance	$\beta_6 > 0$	Acceptance
P8	Joint ETA vs. accuracy-only	ETA terms jointly ≠ 0	Both

Figures

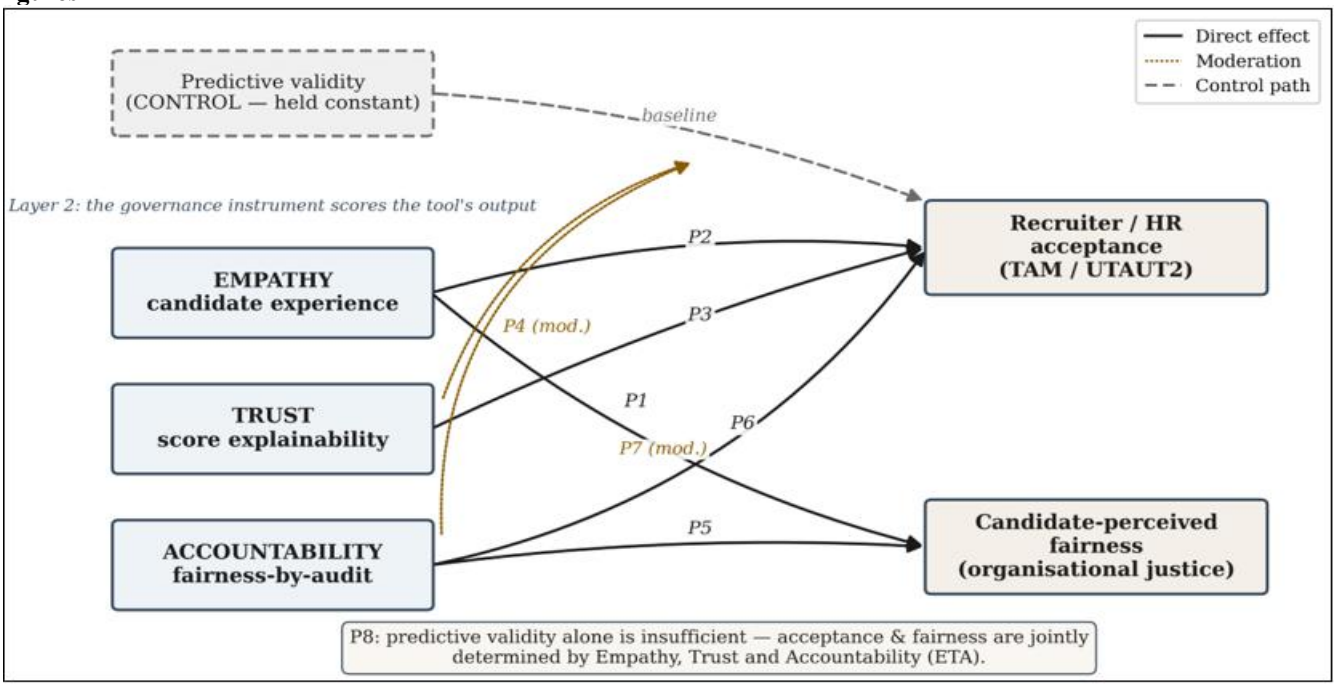

Figure 1

Figure 1: The ETA audit model. Three human-centred axes predict recruiter acceptance and candidate-perceived fairness; accuracy enters as a fixed control; Trust and Accountability moderate the accuracy–acceptance path (P4, P7). Solid = direct effect, dotted = moderation, dashed = control.

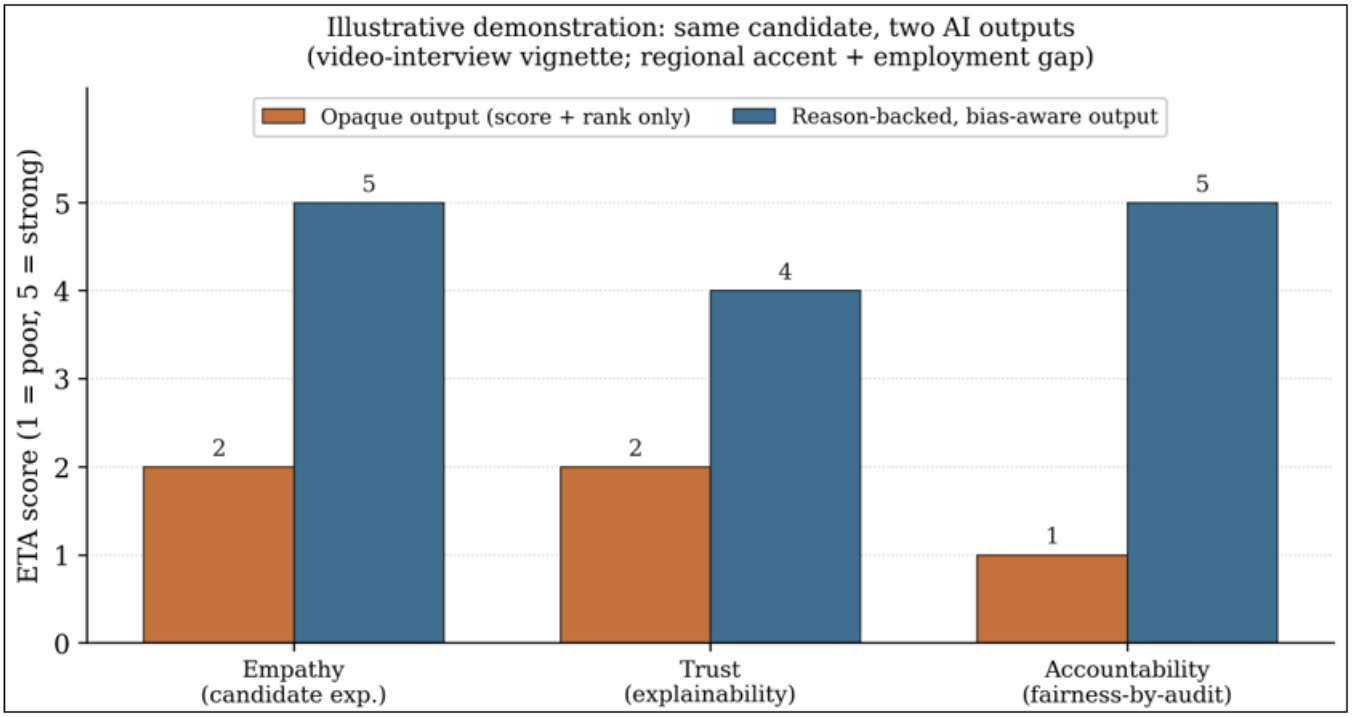

Figure 2

Figure 2: Illustrative module output. For one candidate and one fixed accuracy value, an opaque output and a reason-bearing, bias-aware output receive very different ETA scores. Values illustrate discrimination, not empirical results.