

Counsellor review of artificial intelligence recommendations improves feasibility adjusted career fit and reduces socioeconomic disparity

Karan Gupta

miec99@gmail.com

Karan Gupta Consulting

Research Article

Keywords: artificial intelligence, career guidance, school-to-work transition, future of work, fairness, human-in-the-loop, capability approach, student agency

Posted Date: August 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10516581/v1>

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

[Read Full License](#)

Additional Declarations: Competing interest reported. The author is a practising international education counsellor (Karan Gupta Consulting, Mumbai) and serves as Managing Director, India & South Asia, for IE University. No students with whom the author works professionally participated in the study, and IE University had no role in the study's design, conduct, funding, analysis or the decision to publish. The author declares no other competing interests.

Abstract

Artificial intelligence increasingly mediates how young people enter the workforce, with recommender systems guiding students' educational and career decisions at the start of their working lives. Such systems are evaluated largely on predictive accuracy and user acceptance, leaving open a question central to an inclusive future of work: who benefits, across socioeconomic positions, when AI recommends careers? Drawing on Social Cognitive Career Theory and the Capability Approach, this two-study investigation examines feasibility-adjusted career fit—alignment between a recommendation and a student's interests and abilities, adjusted for their practical capacity to pursue it. Study 1, a transparent policy simulation (20,000 profiles; five guidance policies; three labour-market scenarios), shows that accuracy-optimised recommendation maximises unadjusted fit while widening socioeconomic gaps and assigning infeasible pathways to half of bottom-quintile students; a demographic-parity constraint equalises access but not outcomes; feasibility-informed selection—by constraint-aware AI or counsellor review—produces the highest simulated feasibility-adjusted fit and fewest infeasible recommendations, but sharply reduces poorer students' exposure to high-opportunity pathways unless paired with capability-expanding support. Study 2, a randomised experiment with 987 Indian secondary and university students, finds counsellor-reviewed AI guidance produced the highest perceived feasibility-adjusted fit ($d = 0.52$ versus AI-only) and the smallest socioeconomic disparity, while AI-only guidance depressed reported agency and uncertainty awareness and elevated algorithmic deference ($d = 1.20$); explanation partially restored calibration. AI-assisted guidance should be evaluated as infrastructure for a just school-to-work transition: by feasibility, equity, calibrated uncertainty and preserved agency, not predictive fit alone.

1. Introduction

1.1 Artificial intelligence and the changing gateway to work

Artificial intelligence is transforming not only the content of work but the process by which people find their way into it. Global assessments of the future of work anticipate large-scale reallocation of tasks and skills over the coming decade [1], continuing a debate about computerisation and employment that has run since early estimates of automation exposure [2] and their more sceptical re-examinations [3]. Most of this literature concerns workers already in employment. Yet the same algorithmic infrastructure is moving rapidly upstream, to the point where labour-market trajectories begin: systematic reviews document a fast-growing body of AI applications in education and guidance [4, 5], national case studies describe AI-based career planning entering school systems in the Global South [6], and a recent synthesis of AI in university career counselling reports accelerating institutional adoption alongside thin evidence on student outcomes [7].

This upstream shift matters for the workforce debate because career guidance is one of the few policy instruments that demonstrably shapes school-to-work transitions. Teenagers' career readiness—exploration, alignment between ambitions and educational plans, and contact with the labour market—

predicts later employment and wage outcomes [8]. Guidance is also deeply implicated in social justice: access to high-quality advice is stratified by class, caste, gender and geography, and guidance systems can either compensate for or compound those inequalities [9]. School-to-work transition regimes differ across countries, but in most systems the costs of early mis-navigation are borne disproportionately by young people with the fewest resources [10]. In much of the Global South the scarcity of trained counsellors is acute; in India, career guidance has historically been concentrated among urban and relatively advantaged students, while most young people navigate high-stakes choices with informal advice [11]. India's National Education Policy 2020 explicitly calls for expanded guidance and counselling capacity [12]—a gap that AI systems are now being marketed to fill, at a scale and speed that outpaces evaluation.

This study accordingly treats career-guidance AI as part of the infrastructure of workforce transition rather than merely an educational technology. Its design can influence which young people perceive emerging occupations as genuinely available to them, whether existing inequalities are reproduced at the entry point to work, and whether AI augments or displaces the human judgement required for an inclusive transition. The question it asks—one that current evaluation practice cannot answer—is: when AI recommends careers to students, who benefits across socioeconomic positions, and does adding explanation or human oversight change the answer?

1.2 What career-guidance AI optimises—and what it leaves out

The technical literature on student-facing career recommendation has matured quickly. A systematic review of machine-learning approaches to student career prediction finds a field organised almost entirely around classification accuracy: models are compared on how well they predict a student's stream, degree or occupation from grades, aptitude scores and interest inventories [13]. Work on explainable classifiers for career counselling extends this by making feature contributions visible, but retains predictive performance as the criterion of merit [14]. Meanwhile, the counselling-psychology tradition evaluates guidance through psychosocial outcomes—decision certainty, career maturity, self-efficacy—and meta-analytic evidence shows such interventions work, with effects driven by identifiable ingredients such as individualised interpretation and attention to support-building [15].

Three blind spots follow from this division of labour. First, feasibility: a recommendation can be an excellent match to interests and measured ability yet practically unpursuable—too costly, too distant, or dependent on family circumstances the model never sees. Information-access research has argued that recommender systems must be evaluated for the distribution of exposure and utility across stakeholder groups, not only average relevance [16]; career guidance is a paradigm case, because the “utility” of a recommendation is realised only if the pathway can actually be travelled. Second, distribution: an average improvement in recommendation quality is compatible with widening gaps between advantaged and disadvantaged students. Third, agency: guidance is not allocation. Treating a career recommendation as a prediction problem abstracts away the social and developmental context in which

the advice lands—precisely the abstraction error that sociotechnical critiques of fair machine learning warn against [17]. These blind spots motivate the outcome at the centre of this study, feasibility-adjusted career fit, and the theoretical framework to which we now turn.

1.3 Theoretical framework: social cognition and capability

Two complementary theories structure the investigation. Social Cognitive Career Theory (SCCT) explains how career interests, goals and choices develop through self-efficacy beliefs, outcome expectations and—critically—perceived barriers and supports [18]. In its career self-management extension, SCCT treats guidance episodes as inputs to adaptive career behaviour: advice changes what students believe is possible for them, not merely what they know [19]. SCCT thus predicts that the form of guidance matters: advice that ignores a student’s constraints may depress self-efficacy or be discounted as unrealistic, while advice that engages constraints can convert aspiration into actionable goals. Older but enduring insights about circumscription and compromise add that young people prune their aspirational space early, on the basis of perceived accessibility as much as fit [20]; guidance systems that misjudge accessibility therefore intervene in a process already biased toward premature foreclosure for disadvantaged youth. The psychology-of-working tradition presses the point further: for most people, work choices are made under economic constraint, and theories that assume a free choosing agent misdescribe precisely the students most in need of guidance [21, 22].

The Capability Approach supplies the normative complement. Sen’s core distinction between resources and capabilities—what a person is actually able to be and do—reframes what a career recommendation is for [23]. A recommendation is valuable insofar as it expands a student’s real freedom to pursue a life they have reason to value, not insofar as it is statistically well-predicted [23, 24]. Operationalising capability requires attention to conversion factors: the personal, social and environmental conditions that determine whether formal opportunities convert into real ones [25]. Applied to education and aspiration, the approach highlights how schooling systems can equalise formal access while leaving capability gaps intact [26]. From this perspective, the pathologies of accuracy-optimised guidance are predictable: a system that recommends the “best-fitting” career without modelling conversion factors will systematically overestimate the opportunity sets of constrained students—and any fairness correction that equalises recommendations without expanding capabilities will redistribute labels rather than freedoms.

Together the two theories yield the study’s central construct and its causal logic. Feasibility-adjusted career fit (FAF) is defined as the alignment between a recommended career and a student’s interests and abilities, adjusted for the student’s practical capacity—financial, geographical, informational and social—to pursue the required pathway. FAF is a capability-informed extension of the fit constructs that dominate vocational psychology, and an equity-relevant alternative to the accuracy metrics that dominate recommender evaluation. The hypothesised chain runs: guidance design configuration → recognition of the student’s barriers and supports → calibration of uncertainty and preservation of agency → feasibility-adjusted response to the recommendation → distributional consequences across socioeconomic positions.

1.4 Fairness, explanation and oversight: three candidate remedies

If accuracy-optimised guidance under-serves constrained students, three families of remedy present themselves.

Fairness-aware algorithms. The algorithmic fairness literature documents how systems trained on historical data reproduce disparate impact [27], surveys a proliferation of formal fairness criteria [28], and supplies cautionary empirical cases—most famously a widely used health algorithm whose choice of proxy label systematically disadvantaged Black patients at equal need [29]. Education-specific reviews show the same mechanisms operating on learners [30, 31]. Yet formal criteria embody different and partly incompatible notions of equity [32, 33], and imposing a parity constraint can satisfy the metric while leaving welfare-relevant outcomes untouched or worse [34]. Whether fairness constraints help students—rather than fairness dashboards—is an empirical question that requires modelling what happens after the recommendation is issued. Study 1 examines one widely used criterion, demographic parity of access to high-opportunity recommendations; it is a test of that criterion, not of algorithmic fairness generally.

Explanation. Explainable AI (XAI) research argues that explanations should be built for human reasoners, drawing on the social sciences of explanation [35], and provides a taxonomy of methods for making black boxes interrogable [36]. For guidance, the behavioural question is whether explanations calibrate reliance. People discount algorithmic advice after seeing it err [37] yet in other settings weight algorithmic advice more heavily than human advice [38]—a combination that makes uncalibrated student populations vulnerable in both directions. The automation literature has long distinguished use, misuse and disuse of automated aids [39], with trust calibration as the central design goal [40]. Experimental work shows that explanations can increase unwarranted reliance unless they provoke analytic engagement [41], that human-algorithm combinations can underperform either alone when reliance is poorly calibrated [42, 43], and a meta-analysis of 106 experiments finds human–AI combination is not beneficial on average, though it tends to help when humans hold complementary information [44]—exactly the situation of a counsellor who knows a student’s circumstances.

Human oversight. Ethics guidelines converge on human oversight as a governance principle [45], and UNESCO’s global recommendation makes human determination central to educational uses of AI [46]. The EU Artificial Intelligence Act classifies AI systems used to determine access, admission or assignment to educational and vocational institutions as high-risk, attracting mandatory oversight requirements [47]; advisory career recommenders do not automatically fall within that category, but the direction of travel in governance is clear. What governance frameworks presuppose, however, evidence has not yet supplied: whether counsellor review of AI recommendations actually improves feasibility, reduces disparities and preserves agency—without destroying the scale advantages that motivated AI adoption—has, to our knowledge, not been experimentally examined in career guidance.

1.5 The present research

We combine a transparent policy simulation with a randomised experiment—a design that pairs the internal manipulability of formal modelling [48] with the ecological evidence of a field experiment on real students. Because the two studies operationalise the focal construct differently, we distinguish them throughout: Study 1 measures simulated FAF (a multiplicative index of modelled fit and feasibility), Study 2 measures perceived FAF (students' self-reported fit and feasibility). The constructs are related but not equivalent; Section 3.4 reports sensitivity analyses under both additive and multiplicative specifications.

Study 1 simulates 20,000 student profiles varying in interests, abilities and socioeconomic circumstances, and compares five guidance policies—conventional counsellor heuristics, accuracy-optimised AI, demographic-parity-constrained AI, constraint-aware AI, and human-in-the-loop AI—on mean simulated FAF, socioeconomic FAF gaps, access to high-opportunity careers and infeasible-recommendation rates, under baseline, technology-boom and downturn labour markets. The constraint-aware arm, which gives the algorithm the conversion-factor information the counsellor holds, allows the simulation to separate the value of feasibility information from the value of human judgement per se.

Study 2 randomises 987 secondary and university students across 14 Indian institutions to counsellor-only guidance, AI-only guidance, explainable AI guidance, or AI guidance reviewed by a counsellor, and measures perceived FAF, decision clarity, agency, trust, fairness, uncertainty awareness, algorithmic deference and recommendation acceptance.

Five hypotheses were specified, with perceived FAF designated the primary experimental outcome. H1: AI-assisted guidance improves perceived fit and decision clarity relative to counsellor-only guidance. H2: AI-only guidance reduces students' reported agency and uncertainty awareness and increases algorithmic deference relative to counsellor-only guidance. H3: explanation improves uncertainty awareness and reduces deference relative to AI-only guidance. H4: AI plus counsellor review yields the highest perceived FAF and the smallest socioeconomic disparity in perceived FAF. H5 (Study 1): accuracy-optimised allocation raises average unadjusted fit but widens socioeconomic gaps in simulated FAF, while feasibility-informed selection mitigates both at a modest cost to unadjusted fit. The study design, hypotheses and analysis plan were specified in advance in the research protocol approved by the ethics review board (approval SS/2025/192–3, August 2025), prior to data collection; the study was not otherwise preregistered, and we label analyses as confirmatory only where they follow that protocol.

India provides a strategic research site rather than a merely convenient one: it hosts the world's largest youth cohort, an acute and policy-recognised scarcity of trained counsellors [11, 12], and consequently some of the fastest real-world deployment of AI guidance—conditions under which the equity stakes of design choices are at their most visible. Section 4.3 addresses the generalisability of the findings beyond this setting.

2. Study 1: a transparent policy simulation

2.1 Rationale

Field experiments cannot ethically assign students to receive recommendations known to be infeasible, and cannot observe the counterfactual distribution of recommendations under alternative system designs. Simulation permits exactly this: it makes the assumptions of each guidance policy explicit and auditable, and traces their distributional consequences under controlled variation [48]. The simulation is deliberately simple and fully transparent—every parameter and distribution is documented, with the random seed, in the simulation code provided as supplementary material—so that its results function as a formal argument about design logics, not as an empirical estimate of effect sizes. The bootstrap confidence intervals reported below express sampling uncertainty conditional on the model's assumptions; they do not quantify uncertainty about real-world effects.

2.2 Model

Population. We generated $N = 20,000$ student profiles. Each student has a socioeconomic quintile (uniform over 1–5), a latent ability (normal, mean $0.5 + 0.03 \times (\text{quintile} - 3)$, SD 0.15, truncated to $[0, 1]$), an interest profile over ten career sectors (Dirichlet, concentration 0.8, SES-independent), and two conversion factors that rise with SES plus individual noise: financial capacity and mobility freedom. A proxy achievement measure equals latent ability plus an SES-correlated measurement advantage of 0.06 per quintile step plus noise (SD 0.07), reflecting evidence that recorded achievement embeds schooling quality and preparation differences [30, 31].

Careers. Ten sectors mirror the vignette space of Study 2. Each sector carries a documented opportunity level, pathway cost and mobility requirement (full vectors in the released code); sectors with opportunity ≥ 0.65 are classed as high-opportunity.

Outcomes. True fit combines interest alignment (weight 0.6) and ability–demand match (0.4). Feasibility multiplies affordability (pathway cost against financial capacity) and mobility (requirement against freedom). Simulated FAF is the product of true fit and feasibility for the recommended sector. A recommendation is infeasible if feasibility < 0.40 .

Policies. Table 1 summarises the five policies, which differ only in the information they use and the objective they pursue. Conventional guidance is a counsellor heuristic: a noisy view of fit (SD 0.15, reflecting large caseloads and limited labour-market information) combined with observed constraints. Accuracy-optimised AI selects the sector maximising predicted fit, weighted toward opportunity, using the proxy measure and ignoring feasibility. Demographic-parity-constrained AI adds a constraint equalising the share of high-opportunity recommendations across SES quintiles; it implements one specific fairness criterion, chosen because it is the most common in deployed systems. Constraint-aware AI receives the same conversion-factor information the counsellor observes and directly optimises predicted fit $\times$ feasibility, with no human involvement. Human-in-the-loop AI takes the accuracy

model’s top-five shortlist and lets a counsellor—who observes constraints, with mild judgement noise (SD 0.05)—select the option with the best feasibility-adjusted fit. Comparing the last two isolates whether gains come from feasibility information and objectives, or from human judgement as such.

Scenarios. Policies were evaluated under a baseline labour market, a technology boom (opportunity shifts toward technology-adjacent sectors) and a downturn (opportunity reduced by 0.10 across sectors; cost burdens inflated by 25%). Uncertainty was quantified by 200 bootstrap replications.

Table 1
Guidance policies compared in Study 1.

Policy	Information used	Objective
Conventional guidance	Noisy fit; observed constraints	Balance fit and feasibility heuristically
Accuracy-optimised AI	Proxy-based predicted fit; opportunity	Maximise predicted fit
Demographic-parity-constrained AI	As accuracy AI + SES group shares	Predicted fit s.t. parity of high-opportunity access
Constraint-aware AI	Predicted fit + conversion factors	Maximise predicted fit × feasibility
Human-in-the-loop AI	AI top–5 shortlist + observed constraints	Best feasibility-adjusted option on shortlist

2.3 Results

Table 2 reports the key metrics; Fig. 1 displays mean simulated FAF and the socioeconomic gap by policy and scenario.

Accuracy is not benefit. Accuracy-optimised AI achieved the highest unadjusted fit of any policy (0.92 versus 0.84 for conventional guidance) yet a lower mean simulated FAF (0.721 versus 0.729), because its gains evaporate once feasibility is priced in. The distributional picture is starker: the Q5–Q1 gap in simulated FAF rose from 0.415 under conventional guidance to 0.515, and 51.2% of bottom-quintile students received an infeasible recommendation—against 18.6% under conventional guidance. Constraint-blind optimisation converts a feasibility problem into an apparent talent problem: the system “discovers” that poor students fit demanding careers and routes them there, without modelling whether the pathway can be travelled.

Parity of access is not parity of outcome. The demographic-parity-constrained policy did what it was asked: bottom-quintile access to high-opportunity recommendations rose to 52.5%, close to the top quintile’s 58.6%. But mean simulated FAF (0.717) and the FAF gap (0.523) were essentially unchanged from the accuracy policy, and bottom-quintile infeasibility rose to 53.6%. Equalising the distribution of prestigious labels, without modelling conversion factors, redistributes exposure to infeasibility rather than opportunity—a formal illustration of the capability critique [23, 25] and of warnings that fairness

metrics can be satisfied while welfare is not improved [17, 34]. We emphasise that this result concerns one parity criterion; welfare- or capability-aware fairness constraints would behave differently, and the constraint-aware policy below can be read as one such alternative.

The gain comes from feasibility information, not from humans as such. Constraint-aware AI—the accuracy model given the counsellor’s conversion-factor information and a feasibility-adjusted objective—achieved the highest mean simulated FAF (0.771), statistically indistinguishable from human-in-the-loop review (0.769), with the lowest infeasibility rates of any policy (5.1% for the bottom quintile). Human-in-the-loop review reproduced this profile (7.7% bottom-quintile infeasibility) with slightly more noise. The simulation therefore does not show that humans inherently outperform machines; it shows that whoever holds conversion-factor information and optimises feasibility-adjusted fit—an adequately informed algorithm or a counsellor reviewing an algorithm’s shortlist—dominates constraint-blind optimisation. The case for the human in the loop is institutional rather than computational: in real systems, conversion-factor information is exactly what students rarely enter into forms and what models rarely ingest; the counsellor is, at present, its most reliable sensor, and carries accountability that a feature vector does not.

Feasibility protection breeds aspirational conservatism. Both feasibility-informed policies showed a countervailing cost that we regard as the simulation’s most important result: they steered constrained students away from high-opportunity sectors almost entirely (bottom-quintile access 3.1% under constraint-aware AI and 5.4% under human-in-the-loop review, versus 15.7% under conventional guidance and 49.4% under accuracy AI). Feasibility screening, implemented purely as re-ranking, thus institutionalises conservative expectations—circumscription [20] at system scale, imposed most severely on exactly the students whose aspirations are already most constrained. Under the downturn scenario every policy lost ground (mean FAF–5.3% to–9.3%), with the constraint-blind policies losing most and widening their equity gaps further; feasibility-informed policies were the most resilient but pushed bottom-quintile access to high-opportunity sectors as low as 1.6%. The design implication is developed in Section 4: feasibility-aware guidance must be coupled to capability-expanding responses—scholarship linkage, mobility support, information brokering—so that a detected infeasibility becomes a problem to solve rather than an aspiration to prune.

Table 2

Study 1 results, baseline and downturn scenarios (bootstrap 95% CIs in brackets, conditional on model assumptions; N = 20,000 per cell).

Metric	Conventional	Accuracy AI	Parity-constrained AI	Constraint-aware AI	Human-in-loop
Baseline					
Mean unadjusted fit	0.84	0.92	0.92	0.86	0.87
Mean simulated FAF	0.729 [0.726–0.732]	0.721 [0.718–0.724]	0.717 [0.713–0.720]	0.771 [0.769–0.774]	0.769 [0.766–0.772]
Q5–Q1 FAF gap	0.415 [0.409–0.421]	0.515 [0.509–0.521]	0.523 [0.517–0.530]	0.420 [0.415–0.425]	0.428 [0.423–0.433]
Q1 high-opportunity share	15.7%	49.4%	52.5%	3.1%	5.4%
Q1 infeasible share	18.6%	51.2%	53.6%	5.1%	7.7%
Downturn					
Mean simulated FAF	0.688 [0.686–0.692]	0.655 [0.651–0.659]	0.650 [0.646–0.654]	0.730 [0.727–0.733]	0.727 [0.724–0.730]
Q5–Q1 FAF gap	0.451 [0.445–0.458]	0.564 [0.556–0.569]	0.574 [0.566–0.581]	0.454 [0.450–0.460]	0.462 [0.457–0.466]
Q1 high-opportunity share	13.9%	49.1%	52.3%	1.6%	3.2%

3. Study 2: a randomised experiment with students

3.1 Participants

Participants were recruited through 14 secondary schools and universities across six Indian regions (metropolitan, 34.2%; smaller “Tier-2” cities, 65.8%; government-aided and private institutions); data were collected between October 2025 and March 2026. Of 1,120 students who began the study, 289 were allocated to counsellor-only guidance, 287 to AI-only guidance, 276 to explainable AI guidance and 268 to AI plus counsellor review. A total of 987 met the prespecified inclusion criteria (11.9% excluded for failed attention checks [$n = 60$], completion under four minutes [$n = 42$], severe straightlining or careless-response patterns, or insufficient item completion; categories overlapping; criteria fixed before outcome analysis). Exclusion rates did not differ significantly by condition (8.3%–14.1%; $\chi^2(3) = 5.48$, $p =$

.140) or by socioeconomic quintile ($p = .295$), and an intention-to-treat-style sensitivity analysis retaining all 1,120 starters with any outcome data reproduced the primary result (Section 3.5). The primary sample comprised 554 upper-secondary students (56.1%) and 433 undergraduates (43.9%); mean age 18.4 years ($SD = 1.8$, range 16–22); 53.3% female, 43.8% male, 2.9% non-binary or undisclosed. Self-reported socioeconomic position spanned all five quintiles ($Q1 = 116$, $Q2 = 183$, $Q3 = 252$, $Q4 = 272$, $Q5 = 164$). Sample characteristics are summarised in Table 3.

Table 3
Study 2 sample characteristics by condition (primary sample, $N = 987$).

Characteristic	Counsellor-only	AI-only	Explainable AI	AI + counsellor
n	265	253	237	232
Age, M (SD)	18.4 (1.8)	18.4 (1.8)	18.4 (1.9)	18.4 (1.7)
Female, %	50.2	52.2	51.9	59.5
Upper secondary, %	56.2	55.3	55.7	57.3
SES quintile, M	3.33	3.18	3.01	3.22

3.2 Design and procedure

The experiment used a four-arm between-subjects design. After providing baseline measures (family support, financial and mobility constraints, AI literacy, baseline trust in AI, career decision self-efficacy, career uncertainty, prestige orientation), each participant completed a structured interest-and-circumstances profile. A recommendation was then generated for each participant by a deterministic rule-based matching system implemented in Python (scoring rules, not a trained machine-learning or generative model). The system computed a fit score for each of the ten career sectors from weighted interest–sector similarity (60%) and aptitude/academic-readiness compatibility (40%), applied minimum prerequisite thresholds where relevant, and delivered the recommendation through one of 24 standardised vignette templates (V_01–V_24), each instantiated with the matched sector’s representative career and its associated pathway-cost and opportunity profile (supplementary material S4). All materials were administered in English. Financial capacity, mobility, family circumstances and information access were not used in the initial match and were reserved for counsellor review: the initial system was thus strong on fit but blind to conversion factors, mirroring the fit weighting of Study 1’s latent model and the constraint-blindness of its accuracy-optimised policy. Recommendations were generated prospectively, before participants viewed the experimental materials. Conditions then varied as follows:

1. Counsellor-only guidance: the recommendation was presented as prepared by a qualified career counsellor who had reviewed the participant’s profile.
2. AI-only guidance: the recommendation was presented as generated by an AI system, with no rationale.

3. Explainable AI guidance: as (2), plus a plain-language explanation of the main factors behind the recommendation and an explicit statement of the system's uncertainty. (These two components were bundled by design; their separate effects are not identified.)
4. AI + counsellor review: a qualified career counsellor reviewed the participant's complete interest-and-circumstances profile, the AI-generated recommendation and its rationale, and could endorse the recommendation, qualify it, modify the educational pathway, or replace the recommended career where a genuine feasibility problem emerged; every decision and its reason were recorded. Of the 268 recommendations reviewed, counsellors endorsed 193 (72.0%) unchanged, modified the proposed educational pathway in 56 cases (20.9%), and replaced the recommended career in 19 cases (7.1%); the most frequent reasons for intervention were affordability, geographic mobility and unmet prerequisite requirements. Reviews were conducted by three professional career counsellors (two holding doctorates and one a master's degree, in psychology, counselling or education) with 9–15 years of career-guidance experience, following a written review protocol.

Recommendation content was thus held constant across the counsellor-only, AI-only and explainable-AI conditions; in the reviewed condition it could change where the counsellor judged the AI recommendation infeasible. The reviewed arm is therefore evaluated as a human-reviewed AI guidance package—combining visible human engagement with occasional substantive correction—rather than as a pure provenance manipulation; the discussion returns to this point. Recommendation provenance was experimentally manipulated (participants in each arm were told the recommendation source corresponding to their condition), and all participants were fully debriefed about the manipulation at the end of the study.

Randomisation was computerised within institutions. Manipulation checks confirmed the manipulation registered: perceived AI involvement averaged 6.1 on a 7-point scale in the AI arms versus 1.9 in the counsellor-only arm, and perceived counsellor review 6.1 and 5.9 in the counsellor arms versus 1.9 elsewhere; 90.6% of participants passed the categorical manipulation check. Median completion time was 10.8 minutes. This study was approved by the Zigma Six Ethics Review Board, Mumbai, India. The protocol number for this approval is SS/2025/192–3, granted on 31 August 2025. The Board is registered with the United States Department of Health and Human Services, Office for Human Research Protections (IRB registration number IRB00015192) and holds Federalwide Assurance FWA00036459. All methods were performed in accordance with the United States Federal Policy for the Protection of Human Subjects, the ethical standards of the American Psychological Association, and the Indian Council of Medical Research Ethical Guidelines for Biomedical Research on Human Subjects. The Zigma Six Ethics Review Board certified that the protocol met the requirements of these frameworks. All participants provided informed consent, with guardian consent obtained where required.

3.3 Measures

All primary outcomes were measured on 7-point Likert scales and averaged into scale scores; the full instrument, vignette structure and review protocol are provided in the supplementary material (S4). Internal consistency was acceptable for all scales (Cronbach's $\alpha = .70-.81$; McDonald's $\omega = .70-.82$).

Perceived career fit (4 items, $\alpha = .75$, $\omega = .74$) and perceived feasibility (4 items, $\alpha = .81$, $\omega = .82$; costs, distance, family circumstances, information access) combine into perceived feasibility-adjusted fit, computed in the primary analysis as the mean of the fit and feasibility scores. Decision clarity (4 items, $\alpha = .75$); agency (4 items, $\alpha = .77$; e.g., “the final decision felt like mine”); trust in the guidance (4 items, $\alpha = .76$); perceived fairness (3 items, $\alpha = .70$); uncertainty awareness (3 items, $\alpha = .73$; understanding that the recommendation could be wrong and why); algorithmic deference (3 items, $\alpha = .73$; e.g., willingness to follow the recommendation against one’s own judgement); recommendation acceptance (3 items, $\alpha = .70$).

Measurement structure. Because perceived FAF is the study’s central construct, we verified that fit and feasibility are empirically distinct. Sampling adequacy was good ($KMO = .80$). An exploratory factor analysis of the eight fit and feasibility items yielded two factors with eigenvalues 2.94 and 1.89 (all subsequent < 0.65) and clean simple structure: feasibility items loaded .68–.75 on the first factor and fit items .59–.67 on the second, with no cross-loading above .13. The two scale scores correlated only modestly ($r = .21$), confirming that feasibility is not a redundant facet of fit. We note that self-reported feasibility is capability-informed rather than a direct measure of Senian capability, and that measurement invariance across SES groups and conditions was not formally tested; both points are taken up in the limitations.

3.4 Analytic strategy

Perceived FAF was the prespecified primary outcome; decision clarity, agency, uncertainty awareness and deference were secondary outcomes; the remaining scales and all correlational analyses were exploratory. The confirmatory family comprised the omnibus condition tests and the focused contrasts entailed by H1–H4; Holm correction across this family alters no conclusion (all confirmatory $p < .001$ except where noted). Because participants were nested in 14 institutions, the primary model is a linear mixed model with institution random intercepts (ICC for perceived FAF = .076); one-way ANOVAs with Tukey-corrected contrasts are reported for interpretability and agreed with the mixed models throughout. For the prespecified equity test (condition $\times$ SES interaction), inference with 14 clusters used wild-cluster bootstrap p -values (Rademacher weights, 999 replications) alongside cluster-robust standard errors. Sensitivity analyses re-estimated the primary models with (a) a multiplicative (geometric-mean) FAF specification, and (b) the full allocation sample.

3.5 Results

Perceived feasibility-adjusted fit (H1, H4). Condition affected perceived FAF, $F(3, 983) = 16.30$, $p < .001$, $\eta^2 = .047$ (Fig. 2a); the mixed model confirmed the pattern (AI + counsellor review versus counsellor-only: $b = 0.38$, $SE = 0.05$, $p < .001$). AI + counsellor review ($M = 5.08$, $SD = 0.61$) exceeded every other condition (Tukey-adjusted $p \leq .003$; $d = 0.52$ versus AI-only, $d = 0.59$ versus counsellor-only, $d = 0.34$ versus explainable AI). Explainable AI ($M = 4.87$) exceeded counsellor-only guidance ($p = .036$); AI-only guidance ($M = 4.72$) did not differ from counsellor-only ($M = 4.71$; $p = .996$). H1 was thus supported for decision clarity—all AI arms improved clarity relative to counsellor-only guidance, $F(3, 983) = 22.50$, $p < .001$, $\eta^2 =$

.064—and for perceived fit, $F(3, 983) = 10.49$, $p < .001$, $\eta^2 = .031$, but not for perceived FAF under AI-only guidance: unreviewed AI matched, but did not beat, a counsellor once feasibility entered the outcome. H4's first clause was supported. Both sensitivity analyses reproduced the pattern: geometric-mean FAF, $F(3, 983) = 16.97$, $p < .001$, $\eta^2 = .049$, identical ordering; full allocation sample ($N = 1,120$), $F(3, 1116) = 17.45$, $p < .001$.

Agency, uncertainty awareness and deference (H2, H3). Condition powerfully affected reported agency, $F(3, 982) = 47.31$, $p < .001$, $\eta^2 = .126$ (Fig. 2b). AI-only guidance produced the lowest agency ($M = 4.32$), well below counsellor-only guidance ($M = 4.96$; $d = 0.80$), supporting H2. Counsellor review restored reported agency fully ($M = 5.09$, the highest of any arm; $d = 1.01$ versus AI-only)—students did not experience the added AI as diminishing ownership once a human visibly engaged with their circumstances. Uncertainty awareness showed the largest effect in the study, $F(3, 981) = 100.34$, $p < .001$, $\eta^2 = .235$: AI-only participants were least aware that the recommendation could err ($M = 3.80$), whereas explainable AI ($M = 5.01$) and reviewed AI ($M = 4.94$) exceeded even counsellor-only guidance ($M = 4.49$); the explainable-versus-AI-only contrast was $d = 1.45$, supporting H3. Algorithmic deference mirrored this pattern, $F(3, 972) = 83.57$, $p < .001$, $\eta^2 = .205$: deference spiked under AI-only guidance ($M = 5.02$, versus 3.97 counsellor-only; $d = 1.20$), was substantially reduced by explanation ($M = 4.23$; $d = 0.88$) and fell to the lowest level under counsellor review ($M = 3.87$). In exploratory analysis, deference correlated negatively with uncertainty awareness across the sample ($r = -.29$, $p < .001$), consistent with—though not demonstrative of—the interpretation that undisclosed uncertainty, rather than AI involvement per se, accompanies over-reliance [37, 39, 41].

Trust, fairness and acceptance. Trust, $F(3, 983) = 25.55$, $\eta^2 = .072$, and perceived fairness, $F(3, 983) = 22.28$, $\eta^2 = .064$, were highest in the explainable and reviewed arms. Acceptance differed modestly, $F(3, 974) = 6.81$, $p < .001$, $\eta^2 = .021$: AI-only guidance elicited numerically higher acceptance ($M = 4.53$) than counsellor-only guidance ($M = 4.34$). In an exploratory model, this advantage attenuated to non-significance when deference was included ($b = 0.13$, $p = .093$), while the explainable ($b = 0.24$, $p = .001$) and reviewed ($b = 0.32$, $p < .001$) arms retained their advantages. Because deference is a post-treatment variable, this is an association, not a mediation estimate; it nevertheless cautions against treating raw acceptance as evidence of recommendation quality [16].

Socioeconomic equity (H4). Perceived FAF rose with SES in every condition, but the gradient differed sharply by condition (Fig. 3). Under AI-only guidance the SES slope was steepest ($b = 0.34$ per quintile, cluster-robust $p < .001$), yielding a Q5–Q1 gap of 1.36 scale points; the slope was flattened under counsellor-only (interaction $b = -.11$, $p = .018$), explainable AI ($b = -.12$, $p < .001$) and, most strongly, AI + counsellor review ($b = -.19$, $p < .001$; wild-cluster bootstrap $p = .002$), where the Q5–Q1 gap fell to 0.59. The geometric-mean specification gave near-identical interactions ($-.12, -.14, -.20$). Notably, review compressed the gap from below: bottom-quintile students in the reviewed arm reported the highest perceived FAF of any Q1 group ($M = 4.77$ versus 3.97 under AI-only), while top-quintile students were essentially indifferent across AI arms ($M \approx 5.33\text{--}5.37$). On these perceptual outcomes, human

oversight did not redistribute benefit away from advantaged students; it extended to disadvantaged students a benefit the unreviewed system reserved for the advantaged. H4 was supported in full.

Table 4
Study 2 outcome means (SD) by condition.

Outcome	Counsellor-only	AI-only	Explainable AI	AI + counsellor	F(3, ~983)	η^2
Perceived FAF	4.71 (0.63)	4.72 (0.75)	4.87 (0.63)	5.08 (0.61)	16.30***	.047
Perceived fit	4.77 (0.79)	4.84 (0.78)	4.97 (0.75)	5.13 (0.76)	10.49***	.031
Feasibility	4.65 (0.82)	4.61 (1.16)	4.77 (0.89)	5.03 (0.81)	9.62***	.029
Decision clarity	4.57 (0.85)	4.78 (0.85)	5.00 (0.77)	5.12 (0.80)	22.50***	.064
Agency	4.96 (0.83)	4.32 (0.77)	4.70 (0.74)	5.09 (0.77)	47.31***	.126
Trust	4.54 (0.83)	4.70 (0.79)	5.08 (0.78)	5.02 (0.81)	25.55***	.072
Perceived fairness	4.50 (0.86)	4.41 (0.91)	4.73 (0.87)	5.00 (0.80)	22.28***	.064
Uncertainty awareness	4.49 (0.87)	3.80 (0.82)	5.01 (0.85)	4.94 (0.93)	100.34***	.235
Algorithmic deference	3.97 (0.88)	5.02 (0.86)	4.23 (0.93)	3.87 (0.91)	83.57***	.205
Acceptance	4.34 (0.81)	4.53 (0.83)	4.59 (0.82)	4.65 (0.78)	6.81***	.021
Note. *** $p < .001$. Denominator df vary slightly with item-level missingness (972–983).						

4. Discussion

4.1 Who benefits, and under which design

Across a formal simulation and a randomised experiment, a consistent answer emerges to the question of who benefits across socioeconomic positions: unreviewed, unexplained AI guidance is experienced as most beneficial by the students who need guidance least. In the simulation, accuracy-optimised recommendation raised average predicted fit while widening the socioeconomic gap in simulated FAF and saddling a majority of bottom-quintile students with infeasible advice. In the experiment, AI-only guidance produced the steepest socioeconomic gradient in perceived FAF—top-quintile students responded to algorithmic advice as fully useful, while bottom-quintile students receiving the same

recommendation content experienced it as poorly matched to their real circumstances, in a format that simultaneously depressed their reported agency and their awareness that the advice could be wrong. This conjunction—advice experienced as least feasible by those most inclined to defer to it—is the study’s most troubling finding, and it is precisely the conjunction that accuracy-and-acceptance evaluation cannot detect [13, 16]. We stress the epistemic status of these outcomes: they are immediate, self-reported responses to a single guidance episode, not demonstrated career benefit.

The two remedies tested do different work, and their difference is theoretically informative. Explanation operated on calibration: the explainable arm—which bundled a rationale with an explicit uncertainty statement—produced the study’s largest effect on uncertainty awareness and substantially reduced deference, consistent with accounts of reliance as a trust-calibration problem [40, 41], alongside moderate gains in perceived FAF. (Which component did the work, explanation or uncertainty disclosure, is not identified in this design.) Human review operated on capability-relevant content and engagement: the reviewed arm was a package combining visible human engagement with occasional substantive correction—counsellors endorsed roughly seven in ten AI recommendations unchanged and altered the pathway or career in the remainder, chiefly for affordability, mobility and prerequisite reasons—and this package produced the largest gains in perceived FAF, concentrated among the least advantaged, together with full restoration of reported agency. The present design cannot apportion the package’s effect between visible oversight and substantive correction; a five-arm design separating endorsement-only review from correction-permitted review is the natural follow-up. What the combination does support is the meta-analytic conjecture that human–AI collaboration pays where the human contributes genuinely complementary information [44]: what the counsellor added was not statistical skill but knowledge of conversion factors—exactly the knowledge the model lacked, and exactly the knowledge SCCT and the Capability Approach identify as decisive for constrained students [18, 23, 25].

The simulation sharpens the interpretation. Giving the algorithm the counsellor’s conversion-factor information (constraint-aware AI) reproduced human-in-the-loop performance almost exactly—so the operative ingredient is feasibility information and a feasibility-respecting objective, not human judgement as such. But both feasibility-informed policies exhibited the simulation’s central trade-off: they minimised infeasible advice while steering constrained students almost entirely away from high-opportunity pathways—system-level circumscription [20] that fell hardest on the poorest students, and deepened in a downturn. Feasibility-aware guidance therefore cannot be the whole answer: implemented as pure re-ranking, it risks institutionalising conservative expectations for exactly the students whose aspirations are already most constrained. Oversight should instead convert detected infeasibility into referrals for capability-expanding support—scholarships, mobility assistance, information brokering—so that the feasibility check widens rather than narrows real freedom [23, 26]. We regard this trade-off, not the superiority of any single policy, as the paper’s most consequential design insight.

4.2 Implications for theory, practice and policy

For vocational theory, the findings argue for elevating feasibility from a background variable to a first-class evaluand: perceived FAF behaved differently from perceived fit throughout (the two scales

correlated only $r = .21$), and it—not fit—carried the equity effects. For the responsible-AI literature, the results supply an educational counterpart to findings that proxy choices drive algorithmic harm [29] and that fairness metrics can be satisfied without welfare gains [34]: the relevant failure in guidance is not the outcome label but the omission of conversion factors from the feature space.

The findings also speak directly to the question of what skills the workforce of tomorrow will need, on two levels. For students, calibrated reliance on algorithmic advice—understanding that a confident recommendation can be wrong, and why—emerges as a foundational competency for navigating an AI-mediated labour market: it was the single largest experimental effect, and the one most clearly manipulable by design. For the guidance profession, the results prefigure a role shift from information provider to overseer and corrector of algorithmic systems: the counsellor’s comparative advantage was precisely the conversion-factor knowledge and accountability that models lack. Preparing both groups is itself an educational task that policy has barely begun.

For practice, the reviewed configuration is promising on implementation grounds: counsellor review of an AI recommendation engages minutes of professional time per student rather than a full manual caseload, which suggests—though the present study did not measure counsellor time or cost, and states this as a hypothesis for implementation research—that oversight could multiply scarce counselling capacity rather than replace it. This matters most in systems such as India’s, where policy ambitions for guidance vastly exceed the counsellor workforce [11, 12]. Disclosure of uncertainty should be treated as a default design obligation: in this experiment, the arm containing an uncertainty statement was the arm in which deference fell most sharply short of the AI-only spike. For policy, the findings give empirical content to the governance direction represented by the EU AI Act’s high-risk classification of educational access systems [47]—while noting that advisory recommenders may sit outside its strict scope—and to UNESCO’s insistence on human determination in educational AI [46]: on this evidence, oversight is not merely a safeguard against error but a mechanism that redistributes the perceived benefits of AI toward the students public guidance systems exist to serve.

4.3 Generalisability beyond a single country

The experiment was conducted in India, and its magnitudes are India-specific. The mechanisms it tests, however, are not country-bound in principle: constraint-blind recommendation, socioeconomically graded feasibility of educational pathways, and student deference to confident algorithmic advice have each been documented separately in a range of contexts [8, 9, 30, 37, 38], and stratified school-to-work transitions characterise most economies in degree rather than kind [10]—though we acknowledge that their joint operation in guidance has been demonstrated here in one national setting. India functions in this design as a lead indicator: counsellor scarcity makes it an early and large-scale adopter of AI guidance [7, 11], so design failures visible there today prefigure those that will arrive with scale elsewhere—much as studies of AI career planning in Bangladesh [6] anticipate dynamics for other counsellor-scarce systems. The simulation reinforces the generality claim at the level of mechanism: it contains nothing India-specific, and it reproduces the experiment’s core distributional pattern from assumptions (constraint-blind optimisation, SES-graded conversion factors) that hold wherever

educational pathways are unequally affordable. The boundary conditions we would expect to matter—and which invite replication—are the strength of the SES–feasibility coupling (weaker in strongly subsidised systems), baseline trust in institutions and in AI, the professional status of counsellors, and the depth of digital inequality in access to AI tools themselves [49].

4.4 Limitations

Several limitations qualify the findings. First, Study 2's outcomes are perceptual and immediate; they capture how students responded to guidance, not whether guidance improved their careers. An alternative reading of the reviewed arm's effect is that visible human engagement raised students' confidence in the recommendation, which inflated reported feasibility, rather than the recommendation becoming more feasible in fact; the recorded counsellor corrections make a pure confidence account unlikely to be complete, but disentangling recommendation change, interpretation change and trust change requires behavioural outcomes (information-seeking, enrolment steps, willingness to invest) and longitudinal follow-up, which are planned. Second, the reviewed arm was a package of provenance, engagement and occasional correction; its components are not separable in this design. Likewise, the explainable arm bundled a rationale with uncertainty disclosure. Third, the simulation's parameters, though transparent and stress-tested across scenarios, are stylised; its bootstrap intervals are conditional on its assumptions, and its contribution is the logic of design comparisons, not point estimates. Fourth, Study 1 tested a single fairness criterion (demographic parity of high-opportunity access); other fairness formulations, including welfare- and capability-aware constraints, deserve direct comparison. Fifth, the sample, while large, multi-site and socioeconomically diverse, is not nationally representative—rural schools are absent, and effects may be larger where connectivity and counselling access are thinner; equally, “who benefits” was examined principally across socioeconomic position, not across gender, caste, rurality or disability. Sixth, with 14 institutional clusters, cluster-level inference has limited resolution despite wild-bootstrap safeguards, and measurement invariance of the scales across SES groups and conditions was not formally established. Cross-national replication with harmonised instruments and behavioural endpoints is the natural next step.

5. Conclusion

AI will be recommending careers to students at scale, in systems rich and poor, long before longitudinal evidence on its effects arrives. This study's contribution is to show—formally and experimentally—that the question such systems are currently built to answer (“how accurately can we predict a fitting career?”) is not the question that determines who benefits from them. Evaluated on accuracy and acceptance, AI-only guidance looks like a success; evaluated on feasibility, equity, calibration and agency, it concentrated its perceived benefits on already-advantaged students while dampening the agency and vigilance of the rest—patterns that explanation and modest counsellor oversight substantially reversed in students' immediate responses. The simulation adds the caution that feasibility-aware designs carry their own equity risk, trading infeasible advice for narrowed aspiration unless coupled to capability-expanding support. AI-assisted career guidance should accordingly be evaluated as part of a just school-

to-work transition: not only by predictive fit, but by whether recommended pathways are feasible, benefits are equitably distributed, uncertainty is understood, and student agency is preserved. Systems that meet those criteria can convert a counsellor shortage from a reason for uncritical automation into an opportunity for equitable augmentation—guidance that widens young people’s real freedoms at the moment they first reach toward the world of work.

Declarations

Funding. The author received no specific funding for this work.

Ethics approval and consent. This study was approved by the Zigma Six Ethics Review Board, Mumbai, India. The protocol number for this approval is SS/2025/192-3, granted on 31 August 2025. The Board is registered with the United States Department of Health and Human Services, Office for Human Research Protections (IRB registration number IRB00015192) and holds Federalwide Assurance FWA00036459. All methods were performed in accordance with the United States Federal Policy for the Protection of Human Subjects, the ethical standards of the American Psychological Association, and the Indian Council of Medical Research Ethical Guidelines for Biomedical Research on Human Subjects. The Zigma Six Ethics Review Board certified that the protocol met the requirements of these frameworks. All participants provided informed consent; guardian consent was obtained for participants under 18. Recommendation provenance was experimentally manipulated with ethics-board approval, and all participants were fully debriefed.

Data availability. The complete simulation code (including all parameters and random seeds) is provided as supplementary material. The de-identified experimental dataset, the counsellor review log summary and the analysis scripts are available from the corresponding author on reasonable request. The research questions, hypotheses and analysis plan were registered with the ethics review board as part of the approved protocol (SS/2025/192-3); the study was not separately preregistered in a public registry.

Competing interests. The author is a practising international education counsellor (Karan Gupta Consulting, Mumbai) and serves as Managing Director, India & South Asia, for IE University. No students with whom the author works professionally participated in the study, and IE University had no role in the study’s design, conduct, funding, analysis or the decision to publish. The author declares no other competing interests.

Use of artificial intelligence. A generative AI assistant (Claude, Anthropic) was used under the author’s direction to assist with drafting and editing the manuscript text, implementing the Study 1 simulation and statistical analysis code, and preparing figures and tables. The study’s conception and design, data collection, counsellor-review protocol, and all final analytical and interpretive decisions are the author’s own; the author reviewed and verified all AI-assisted content and takes full responsibility for the integrity of the manuscript.

Author contributions. KG: conceptualisation, methodology, investigation, formal analysis, writing – original draft, writing – review and editing.

References

1. World Economic Forum. The Future of Jobs Report 2025. Geneva: World Economic Forum; 2025.
2. Frey CB, Osborne MA. The future of employment: how susceptible are jobs to computerisation? *Technol Forecast Soc Change*. 2017;114:254–80.
3. Autor DH. Why are there still so many jobs? The history and future of workplace automation. *J Econ Perspect*. 2015;29(3):3–30.
4. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education—where are the educators? *Int J Educ Technol High Educ*. 2019;16:39.
5. Holmes W, Tuomi I. State of the art and practice in AI in education. *Eur J Educ*. 2022;57(4):542–70.
6. Shabur MA. The potential and implications of artificial intelligence in Bangladesh's early career planning education. *Discov Glob Soc*. 2024;2:50.
7. Sarmurzin Y, Azhenov A, Kudysheva A, Ualikhanova B, Saparbekova I, Besbayev A. Implementation of AI in career counselling for university students: a systematic review. *Front Educ*. 2026;11:1787689.
8. Covacevich C, Mann A, Besa F, Diaz J, Santos C. Indicators of teenage career readiness: an analysis of longitudinal data from eight countries. OECD Education Working Papers No. 258. Paris: OECD Publishing; 2021.
9. Hooley T, Sultana RG, Thomsen R, editors. Career guidance for social justice: contesting neoliberalism. New York: Routledge; 2018.
10. Ryan P. The school-to-work transition: a cross-national perspective. *J Econ Lit*. 2001;39(1):34–92.
11. Arulmani G. The cultural preparation process model and career development. In: Arulmani G, Bakshi AJ, Leong FTL, Watts AG, editors. *Handbook of career development: international perspectives*. New York: Springer; 2014. p. 81–103.
12. Government of India. National Education Policy 2020. New Delhi: Ministry of Human Resource Development; 2020.
13. Trujillo F, Pozo M, Suntaxi G. Artificial intelligence in education: a systematic literature review of machine learning approaches in student career prediction. *J Technol Sci Educ*. 2025;15(1):162–85.
14. Guleria P, Sood M. Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling. *Educ Inf Technol*. 2023;28:1081–116.
15. Whiston SC, Li Y, Goodrich Mitts N, Wright L. Effectiveness of career choice interventions: a meta-analytic replication and extension. *J Vocat Behav*. 2017;100:175–84.
16. Ekstrand MD, Das A, Burke R, Diaz F. Fairness in information access systems. *Found Trends Inf Retr*. 2022;16(1–2):1–177.

17. Selbst AD, boyd d, Friedler SA, Venkatasubramanian S, Vertesi J. Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). New York: ACM; 2019. p. 59–68.
18. Lent RW, Brown SD, Hackett G. Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *J Vocat Behav.* 1994;45(1):79–122.
19. Lent RW, Brown SD. Social cognitive model of career self-management: toward a unifying view of adaptive career behavior across the life span. *J Couns Psychol.* 2013;60(4):557–68.
20. Gottfredson LS. Circumscription and compromise: a developmental theory of occupational aspirations. *J Couns Psychol.* 1981;28(6):545–79.
21. Blustein DL. The psychology of working: a new perspective for career development, counseling, and public policy. Mahwah: Lawrence Erlbaum; 2006.
22. Duffy RD, Blustein DL, Diemer MA, Autin KL. The psychology of working theory. *J Couns Psychol.* 2016;63(2):127–48.
23. Sen A. Development as freedom. Oxford: Oxford University Press; 1999.
24. Nussbaum MC. Creating capabilities: the human development approach. Cambridge: Harvard University Press; 2011.
25. Robeyns I. The capability approach: a theoretical survey. *J Hum Dev.* 2005;6(1):93–117.
26. Hart CS. Aspirations, education and social justice: applying Sen and Bourdieu. London: Bloomsbury; 2012.
27. Barocas S, Selbst AD. Big data's disparate impact. *Calif Law Rev.* 2016;104(3):671–732.
28. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv.* 2021;54(6):1–35.
29. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447–53.
30. Baker RS, Hawn A. Algorithmic bias in education. *Int J Artif Intell Educ.* 2022;32:1052–92.
31. Kizilcec RF, Lee H. Algorithmic fairness in education. In: Holmes W, Porayska-Pomsta K, editors. The ethics of artificial intelligence in education. New York: Routledge; 2022. p. 174–202.
32. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12). New York: ACM; 2012. p. 214–26.
33. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems 29 (NeurIPS 2016); 2016. p. 3315–23.
34. Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S. The measure and mismeasure of fairness. *J Mach Learn Res.* 2023;24(312):1–117.
35. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell.* 2019;267:1–38.

36. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138–60.
37. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J Exp Psychol Gen*. 2015;144(1):114–26.
38. Logg JM, Minson JA, Moore DA. Algorithm appreciation: people prefer algorithmic to human judgment. *Organ Behav Hum Decis Process*. 2019;151:90–103.
39. Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. *Hum Factors*. 1997;39(2):230–53.
40. Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors*. 2004;46(1):50–80.
41. Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc ACM Hum-Comput Interact*. 2021;5(CSCW1):188.
42. Green B, Chen Y. The principles and limits of algorithm-in-the-loop decision making. *Proc ACM Hum-Comput Interact*. 2019;3(CSCW):50.
43. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York: ACM; 2021. Article 81.
44. Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nat Hum Behav*. 2024;8:2293–303.
45. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1:389–99.
46. UNESCO. Recommendation on the ethics of artificial intelligence. Paris: UNESCO; 2021.
47. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union*. 2024;L 2024/1689.
48. Epstein JM. Why model? *J Artif Soc Soc Simul*. 2008;11(4):12.
49. van Dijk J. The digital divide. Cambridge: Polity Press; 2020.

Figures

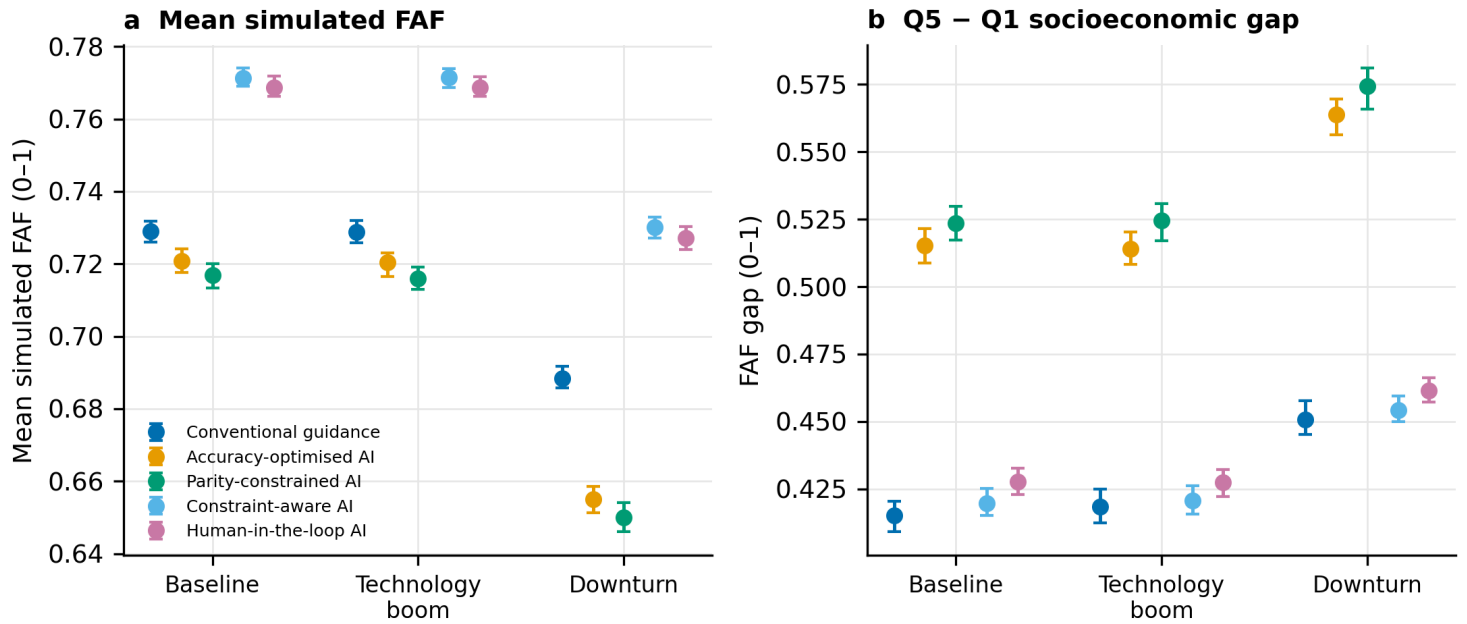


Figure 1

Study 1: (a) mean simulated feasibility-adjusted fit and (b) Q5–Q1 socioeconomic gap by guidance policy across labour-market scenarios. Points show means with bootstrap 95% confidence intervals, conditional on model assumptions. Note the focused y-axes, chosen to resolve between-policy differences; full-scale versions are provided in the supplementary material.

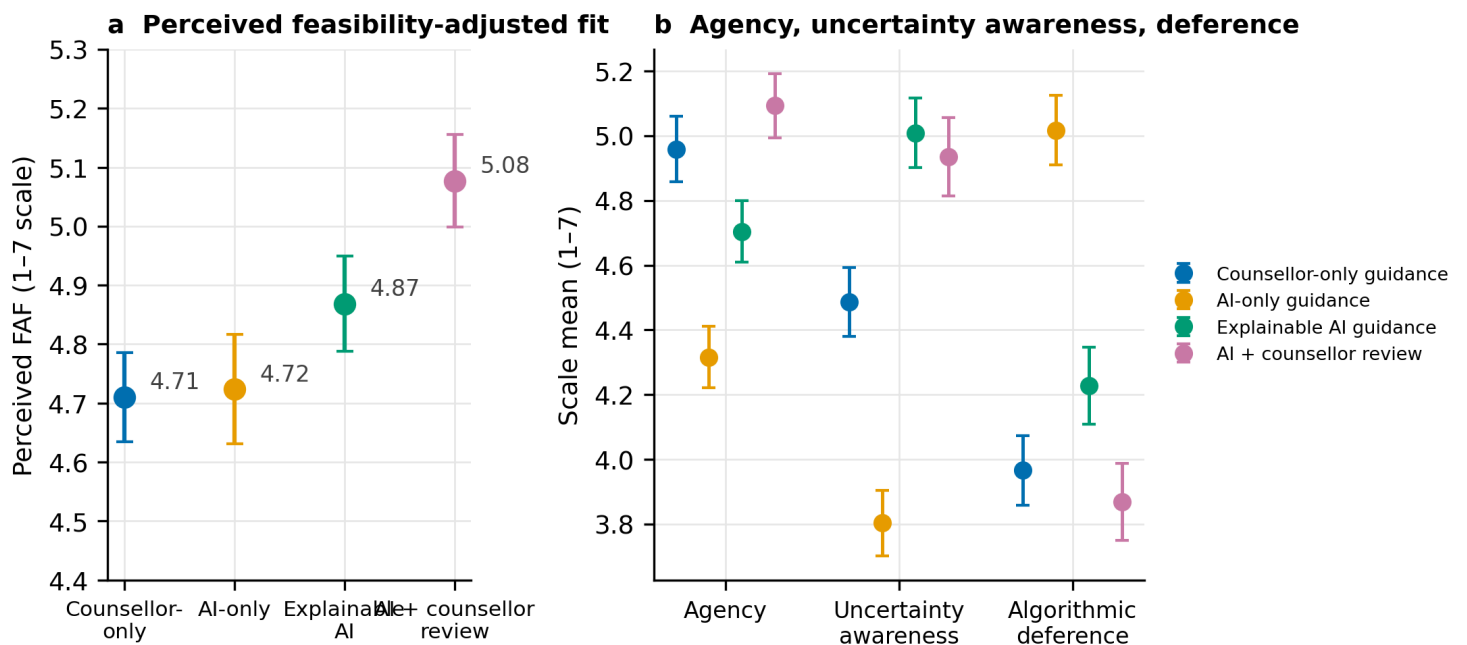


Figure 2

Study 2: (a) perceived feasibility-adjusted fit by condition; (b) reported agency, uncertainty awareness and algorithmic deference by condition. Points show means with 95% confidence intervals. Y-axes are truncated

to resolve between-condition differences on the 1–7 response scales.

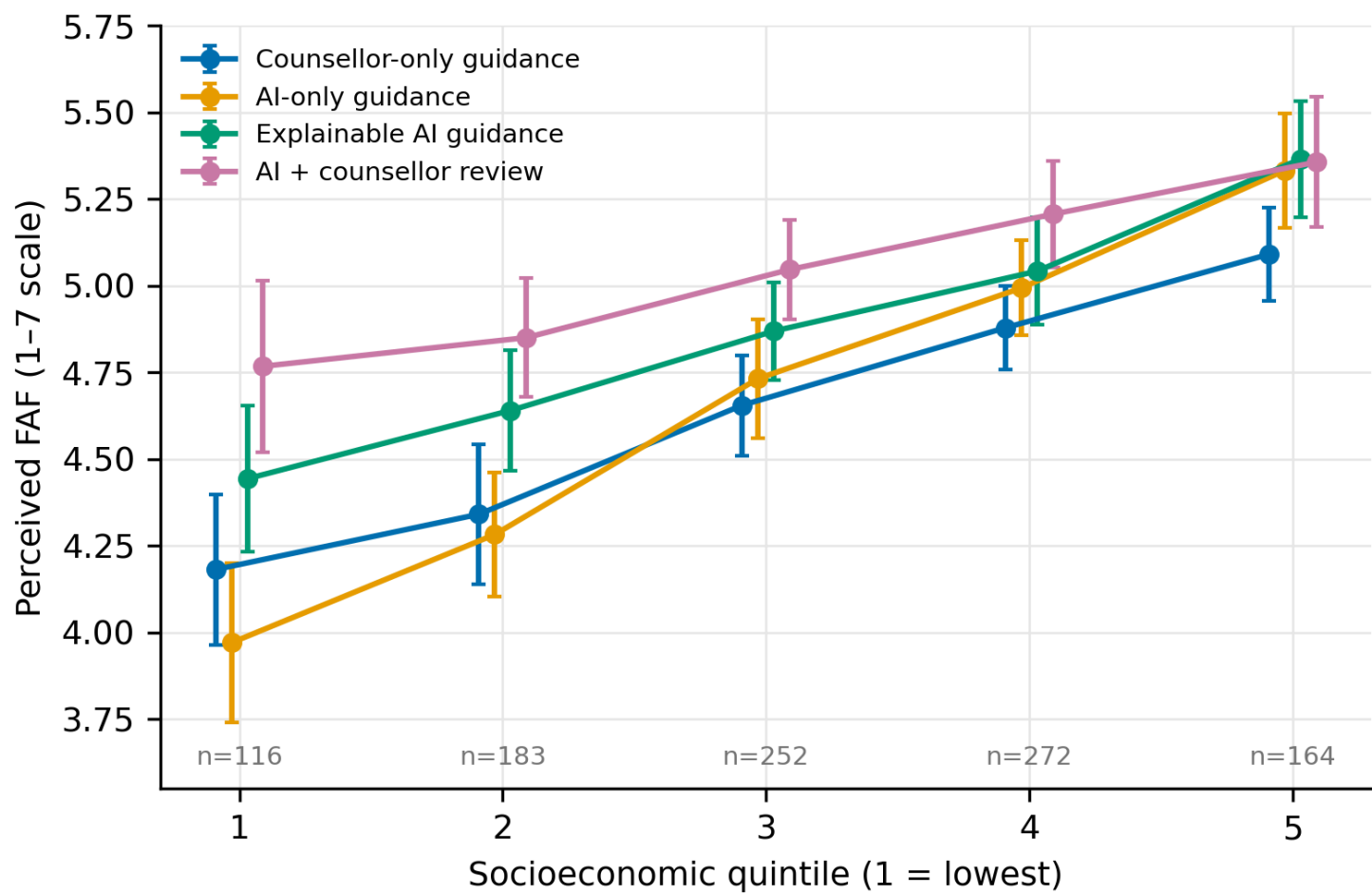


Figure 3

Perceived feasibility-adjusted fit by socioeconomic quintile and condition (raw cell means with 95% confidence intervals; total n per quintile shown at base). The socioeconomic gradient is steepest under AI-only guidance and flattest under AI + counsellor review. Y-axis truncated; response scale 1–7.

Supplementary Files

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

- [SupplementaryMaterial.docx](#)