

Rights, Reasons, and Results: Ethics for AI Hiring in Finance

Qualifying Exam Question 2: Dr. Donald Williams

Christopher Law

11/11/2025

Muskingum University

Department of Organizational Leadership

Financial institutions are increasingly turning to algorithmic screening, ranking, and assessment to cope with surging applicant volumes and to sharpen their competitive edge in the market for scarce skills. This shift raises a question that guides everything that follows: how should leaders evaluate the moral legitimacy of AI-mediated hiring with respect to transparency, accountability, and the mitigation of unjust bias? Evidence from large and mid-sized banks shows that algorithmic tools already shape first-pass filtering, résumé ranking, and standardized assessments that compress time to hire and promise procedural consistency (Bartlett, Morse, Stanton, & Wallace, 2022; Fuster, Goldsmith-Pinkham, Ramadorai, & Walther, 2022; Berg, Burg, Gombović, & Puri, 2020). Interpreting these findings through a normative lens shows why technical promise alone is insufficient.

In finance, where public trust functions as reputational capital, legitimacy depends as much on the duties owed to applicants as on aggregate outcomes. That observation justifies evaluating hiring systems first by the claims they must respect and only then by the benefits they may deliver. Because AI hiring is a socio-technical process, outcomes emerge from the coupling of models with data pipelines, user interfaces, recruiter workflows, and vendor contracts, as well as cultural and socioeconomic contexts, rather than solely from the algorithms themselves. Accordingly, this paper treats legitimacy as a property of the entire system of practice (encompassing people, process, and technology), rather than a single model artifact.

Deontology provides the baseline grammar for a rights-first evaluation. In its classical Kantian form, deontology assesses conduct by its conformity to duties and by the respect it accords to persons, which requires acting only on maxims that could be willed as universal law and treating persons as ends, not as mere means (Kant, 1785/2011). A plural-duty tradition articulated by Ross adds practical traction by recognizing fidelity, reparation, justice,

beneficence, non-maleficence, and self-improvement as prima facie duties whose all-things-considered resolution calls for judgment in context (Ross, 1930/2002). In rights-sensitive domains, such as employment, deontology functions as a set of side constraints that determine what cannot be done to people, even when a tempting gain presents itself.

Consequentialism approaches the same terrain by evaluating actions based on their effects on overall welfare across stakeholders and over time. In organizational life, this orientation translates into risk identification, cost–benefit analysis, scenario testing, and ongoing monitoring that map, measure, and manage socio-technical risk (Mill, 1863; National Institute of Standards and Technology [NIST], 2023). Consequentialist tools reinforce a discipline of evidence: if a compliant design does not improve quality-adjusted time to fill, retention, performance, and diversity, or if it externalizes harms to applicants, employees, or the environment, its advocates bear the burden of proof. Therefore, duties set red-lines for what is permissible, and within that space, welfare evidence guides practical selection among compliant designs.

Placed alongside governing law and supervisory expectations, a deontological lens should lead in employment. In the United States, Title VII’s disparate-impact doctrine, as well as the Equal Employment Opportunity Commission and Department of Justice’s technical assistance on AI and disability, require non-discrimination, reasonable accommodation, and defensible job-relatedness in screening tools. New York City’s Automated Employment Decision Tools regime requires candidates to be notified and undergo a documented bias audit for covered uses. The European Union’s Artificial Intelligence Act classifies many employment systems as high-risk, mandating risk management, documentation, and post-market monitoring (European Union, 2024; New York City Department of Consumer and Worker Protection, 2023; U.S.

Department of Justice & U.S. Equal Employment Opportunity Commission, 2022). Credit sector rules under ECOA and Regulation B require explanation specificity for lending decisions and provide a helpful analogy for case-level reasons in hiring. However, they are not binding in employment settings (an analogy only, not a source of legal obligation). In this landscape, a rights-first sequence respects binding duties, while a welfare-oriented cross-check verifies that compliant designs deliver the benefits they claim.

AI-mediated hiring in finance is a pipeline rather than a point decision. Data enters through job boards, referrals, and internal systems. Vendors or internal teams engineer features, models, rank, and filter, and outputs are integrated into recruiter workflows that lead to interviews and offers. At each transformation, distinct risks appear. Feature construction can import proxies for protected characteristics. Ranking functions can amplify minor score differences into significant opportunity gaps at short-list boundaries. Workflow integrations can encourage uncritical acceptance of model output at the expense of reflective judgment. These records require versioning, owners, escalation paths, and audit trails to have probative value in practice (Mitchell et al., 2019; Gebru et al., 2021; FinRegLab, 2023). Governance and accountability become diffuse and cannot be located at a single checkpoint. It must assign human roles and checkpoints across the pipeline, from data-collection notices and consent records to Model Cards, Datasheets, explanation libraries, decision logs, and appeal files, so that an external reviewer can reconstruct what happened and why for any given decision.

The institutions that build or buy these tools do so for compelling operational reasons. Applicant volumes have surged, especially at the entry level, and competition for scarce skills has intensified. Documented effects of automation include faster triage and fewer idiosyncratic first-pass decisions. However, adjacent finance literatures warn that the same predictive strength

that improves signal can also reproduce disparities through proxies if governance is weak (Fuster et al., 2022; Bartlett et al., 2022; Berg et al., 2020). In a rights-sensitive domain, that evidence argues against vendor-first pragmatism that treats outcomes as sufficient. Instead, it supports a theory-led approach that asks whether applicants are respected as claim-holders with rights to notice, explanation, and appeal; whether responsibilities are visible and enforceable inside the firm and along the vendor chain; and whether effects on groups are monitored and mitigated over time. The ordering is deliberate: a deontic floor for legitimacy first, followed by a consequentialist verification that the floor holds in the world.

There are already several ethics frameworks available to help practitioners understand how these tools align with this sequence. Some approaches function primarily as duty frameworks and prioritize rights to notice, explanation, non-discrimination, and role responsibility as obligations owed to affected persons. Others operate as outcome frameworks that emphasize the measurement of aggregate effects, lifecycle risk, and environmental or organizational externalities. A third group provides governance scaffolding that blends minimum rights with continuous monitoring and iterative improvement. The Ethic Pocketknife's attention to transparency, accountability, and empathy is duty-leaning; the Power of Ethics' structured reasoning is also duty-leaning as it centers principled deliberation about what is owed to whom; EMMA's management-level integration of ethical and environmental dimensions belongs to hybrid governance; and stakeholder models that specify roles, responsibilities, and assurance routines likewise occupy the hybrid space because they combine minimum constraints with monitoring and adaptation (Bühler, 2021; Liautaud & Sweetingham, 2022; Brendel, Mirbabaie, & Lembcke, 2021; Lau, 2021). Read through a socio-technical lens, hybrid governance frameworks are strongest when they pair minimum rights with operator training, workflow

controls, and interface-level safeguards that shape how humans actually use model outputs (NIST, 2023; ISO/IEC, 2023).

Across the leading frameworks, two gaps persist in employment contexts. Most assume adoption rather than guiding the predicate question of whether a given AI tool should be used at all in a particular setting. Few operationalize rights as specific artifacts, thresholds, and stop-rules that bind deployment. These gaps manifest as three concrete deficits that the present blueprint addresses. First, at the feature level, there is often no explicit job-relatedness test, no documented consideration of less discriminatory alternatives, and no threshold for what counts as acceptable disparity. Second, the applicant-facing duties are weak or undefined. Notice language lacks specificity, and explanation practices do not ensure stability across similarly situated cases; appeals lack service levels that make redress real. Third, vendor accountability is limited. Documentation is partial, change-control is informal, incident response is slow, and environmental externalities at inference are rarely measured or managed.

The duties only matter if they can be shown in practice; the next phase translates them into auditable tests. Respect for persons requires truthful, plain-language notice of automated processing, disclosure of data categories and inferences used, and a meaningful path to human review. Employment doctrine adds two specific implications. Any AI-mediated assessment should be paired with a non-automated pathway upon request (U.S. Department of Justice & U.S. Equal Employment Opportunity Commission, 2022). The artifact for this duty is a candidate-facing notice and consent record, posted when an application is initiated and again at the point at which a tool will materially influence ranking. The threshold is clarity and timing as verified by usability testing and a signed consent record tied to the application ID. The owner is Human Resources, and the escalation path is to legal and model-risk when complaints or

usability tests reveal confusion. In socio-technical terms, respectful design also requires consent-based UX that avoids dark-pattern nudges, accessibility accommodations (e.g., screen-reader-friendly notices), and documented alternatives for candidates whose abilities interact poorly with the interface. These are design decisions, not afterthoughts, and should be validated with usability testing, not just legal review (NIST, 2023; HLEG, 2019).

AI-mediated hiring is a socio-technical system, where outcomes emerge from the interactions among data, models, interfaces, policies, and human judgments throughout the hiring workflow. Contemporary guidance treats AI risk in precisely these terms, emphasizing that governance must cover people, process, as well as model behavior (NIST, 2023; European Commission High-Level Expert Group on AI, 2019). On that view, moral legitimacy depends not only on what the model predicts but on how organizations design notices, set thresholds, train reviewers, track auditable artifacts, and respond to drift. The artifact is a feature rationality catalog that documents the job relevance of each feature, providing evidence and flagging high-risk proxies. The threshold is a pre-committed disparity band for error-rate parity and calibration within groups, with equalized odds where qualification rates permit. The owner is the model risk for validation, and HR co-owns the job analysis. The escalation path is to a cross-functional committee that can halt deployment if parity bands are breached. A documented alternatives analysis for each mitigation choice operationalizes that standard in model development and review (NIST, 2022; Mitchell et al., 2019). Where results show statistically significant disparities, the employer must demonstrate that the practice is business-necessary and consider less discriminatory alternatives. A socio-technical justice check also examines channel effects: shifts in sourcing (e.g., job boards, internal referrals, campus pipelines) alter the applicant distribution the model observes and can create performative feedback, where optimization for

speed reshapes who applies. Representational harms (e.g., stereotyping in textual features) can coexist with allocative harms (e.g., fewer interviews). Testing should therefore cover pre-model decisions (where applicants come from), in-model proxies, and post-model human interventions, with parity bands applied at each stage (shortlist thresholds, interview invitations, and offers) (NIST, 2022; Raji et al., 2020).

Evaluating the actual processes involved with audit trails and documentation should not stop at the legal limit if the goal is to avoid reputational risk. Deontology provides a basis for evaluating the “why” of the process and keeping organizational and ethical values in mind for all parties involved, including the employee, the organization, and the hiring managers. Most existing ethical frameworks are either too vague to be relevant to individual processes or overly specific to their use case, bogging down in excessive detail. Crucially, governance frameworks earn their keep when they explicitly span socio-technical elements, including people, processes, and platforms — rather than treating models in isolation (NIST, 2023; European Commission High-Level Expert Group on AI, 2019). Using core concepts of Deontology, complex and diffuse processes can be distilled into company values that management, vendors, and employees can use to check their processes against reality, develop ethical arguments to change existing unethical practices, and evaluate the implementation goals of new AI decision-making tools.

Concepts of justice forbid unjust discrimination and demand rational job-relatedness for features used in ranking and screening. Independent audits and empirical studies show how models trained on historical patterns can encode structural disadvantages that reappear as seemingly elegant predictors. Fairness evaluation, therefore, must attend to proxies and group-level error behavior, not protected attributes alone (Fuster et al., 2022; Das, Stanton, & Sundaram, 2023). Calibration within groups ensures that probabilities have the same meaning

across different populations. Error rate parity guards against systematic over- or undersampling. Equalized odds are appropriate where qualification rates are not sharply divergent. In ranked screening, top-K parity checks at shortlist boundaries and threshold stability checks across subpopulations are crucial. Explanation stability deserves equivalent rigor. Small, job-irrelevant perturbations should not yield different primary reasons, and candidate-facing reasons should align with internal contribution analyses for the same decision. Where post-hoc techniques such as SHAP or LIME are used, banks should document faithfulness tests that compare post-hoc attributions to model-internal signals on held-out data (NIST, 2022; FinRegLab, 2023; Bank for International Settlements, Financial Stability Institute [BIS-FSI], 2024). The duty becomes empirical in the right way when these tests are pre-specified, executed, and reported to owners and management with escalation procedures in place.

Truthfulness requires accurate, specific, and comprehensible reasons for decisions that materially disadvantage applicants. In finance, explanation is not a cosmetic add-on; it is a control function for supervision and internal model-risk governance (FinRegLab, 2023). While ECOA's adverse-action notice regime is specific to credit, employment regimes increasingly require notice and transparency, and prohibit discriminatory screening practices; some jurisdictions already mandate candidate notice or bias audits in specified contexts (New York City Department of Consumer and Worker Protection, 2023). The artifacts include a candidate-facing reason library and an internal explanation registry that maps external reason codes to features and logic, with version control. The thresholds are explained by stability across adjacent cases and alignment of candidate-facing reasons with internal contribution analyses for the same decision. Model risk owns the stability tests, recruiting owns communication, and appeals have published service levels. A reasonable service level would be to acknowledge an appeal within

two business days and provide a reasoned determination within ten business days, allowing for the submission of supplemental information and the request for an independent review.

Explanation is only truthful in use if humans can comprehend and act on it. In addition to stability tests, require explanation usability studies (short comprehension checks with recruiters and applicants) and track whether reasons change downstream decisions in appeals. Treat “readability and actionability” as first-class metrics in model-risk validation (FinRegLab, 2023; NIST, 2023).

Accountability requires visible role assignments, enforceable responsibility for decisions that affect persons, and credible vendor governance. The artifacts include a RACI (Responsible, Accountable, Consulted, Informed) model for the hiring toolchain, a signed vendor obligations matrix, and incident response runbooks. Vendor agreements should require Model Cards and Datasheets for each release, secure data-room access for auditors, subgroup performance dashboards built to agreed-upon group definitions, explanations of stability results, and pre-notification at least thirty days before material changes. Incident-response clauses should require notice within twenty-four hours of detected threshold breaches, with remediation timelines and rollback procedures specified. The thresholds are fulfilled, deliverables are at the go-live gates, and tested escalation drills are in production. The owners are responsible for contract enforcement, model-risk for validation, and HR and compliance for process adherence. Escalation proceeds from the tool owner to a cross-functional ethics and risk council, which has the authority to pause the process.

Beneficence and non-maleficence require avoiding wrongful exclusion and promoting legitimate goods such as broader access, subject to the constraints already named. Summaries of fairness research in financial economics emphasize that equity and accuracy can be compatible

when proxy features are governed and monitoring detects drift in subgroup error rates (Das et al., 2023; NIST, 2022). The artifacts here are a benefits ledger and a harms ledger that track realized gains and adverse effects. The thresholds are pre-committed nets for quality-adjusted time to fill and subgroup performance that trigger review when deteriorations occur. Owners are responsible for HR analytics related to benefits and model risk for potential harms, with joint escalation if trade-offs arise.

To maintain legitimacy over time, go-live gates and stop rules must be clearly defined and consistently applied. European and U.S. instruments require risk management and post-market monitoring that trigger remediation when thresholds are violated and, in some cases, prohibit manipulative practices outright (European Union, 2024; NIST, 2023). In hiring, institutions should pre-commit to disparity bands with escalation provisions when exceeded, barring adverse inference and covert psychometrics. In rare business necessity cases, institutions should document alternatives considered, mitigation steps taken, and a time-bound reevaluation that includes the option to retire features that persistently harm groups. In situations where explainability fails, stop rules are not abstract ideals; they are switches that a named human authority should feel empowered to use.

Having established the floor of ethical permissibility, the analysis then examines whether compliant designs actually improve outcomes and mitigate externalities. A compact consequentialist cross-check can be standardized as an outcome panel. The panel includes quality-adjusted time to fill, post-hire performance indexed to role complexity, retention at one and two years, diversity composition at shortlist and offer stages, appeal rates and outcomes, applicant experience scores for explanation clarity and perceived fairness, and energy-per-inference. Where the outcome panel shows regressions or widening subgroup

disparities, escalation triggers revalidation or rollback regardless of short-term efficiency gains. This discipline is not adversarial to the deontic floor; it protects the institution from slow drift that erodes the meaning of compliance.

Externalities do not stop at the firm boundary, so responsible programs treat the environmental footprint as part of risk and stewardship. Standards bodies recommend lifecycle energy and carbon accounting for training and inference, and they encourage model-size and architecture choices that minimize resource consumption without compromising purpose (NIST, 2023; ISO/IEC, 2023). In hiring, inference dominates. Finance, HR, or a corresponding operations team should record energy per inference in the risk register, using region-specific grid intensity factors, and should prefer architectures measured below an internal target. Where multiple models are viable, selection should include an explicit trade-off entry that describes the accuracy, fairness, and energy differences, and names the owner of the decision. These steps reduce the chance that efficiency gains for recruiters externalize costs to communities and ecosystems.

Banking supervisors expect model-risk programs to detect and remediate distribution shifts and to define incident response that includes notification and rollback. AI guidance emphasizes that the govern–map–measure–manage cycle must be iterative and socio-technical (Office of the Comptroller of the Currency [OCC], 2021/2025; NIST, 2023). Applied to hiring, this means monitoring subgroup calibration and error rates, validating that explanation quality does not decay, and setting explicit triggers that escalate to human review and temporary suspension when thresholds are crossed. Controls must extend to vendor-supplied models via contract and assurance. Procurement integrates obligations, model-risk validates evidence, and

HR enforces appeal SLAs. A single dashboard displays the state of compliance and outcomes, the owner of each control, and the escalation path when warnings arise.

At a universal bank deploying a résumé-ranking service for long-list triage, the go-live plan begins with job analysis to justify each feature’s relevance, alongside pre-deployment tests for subgroup calibration and error behavior (Raghavan, Barocas, Kleinberg, & Levy, 2020; NIST, 2022). Upon application, candidates receive a clear notice that automated assistance will inform screening, a link describing data categories and logic at a high level, and the right to human review, an approach consistent with emerging employment-specific transparency regimes such as New York City’s AEDT rules (New York City Department of Consumer and Worker Protection, 2023). When the system screens out a candidate, case-level reasons are generated from a version-controlled library, and the usefulness and stability of the explanation are validated as part of quality assurance (FinRegLab, 2023; Bank for International Settlements—Financial Stability Institute, 2024). Appeals operate under published service levels; outcomes and any human overrides are logged to an auditable decision file. Monitoring shows screening time is falling without widening disparities, and when a quarterly review detects a modest decline in calibration for one subgroup, targeted recalibration and a threshold adjustment bring metrics back within pre-committed bands, with change-control artifacts and rollback plans maintained per model-risk guidance (NIST, 2023; Office of the Comptroller of the Currency, 2021/2025). The result illustrates that the duties of transparency, truthful explanation, and avoidance of unjust bias can be effectively demonstrated in practice without compromising efficiency. At the same time, outcome monitoring provides the discipline to correct any drift (Das, Stanton, & Sundaram, 2023).

When a retail bank evaluates asynchronous video interviews that involve AI-evaluation of stress, grammar, and intonation analysis, the program applies a rights-first screen and finds the “respect for persons” test unsatisfied: affect inference lacks a credible job-related rationale, invites proxy discrimination, and risks disparate impact (Raghavan et al., 2020). Employment doctrine and technical assistance on disability further caution against screening practices that disproportionately exclude qualified applicants or that fail to provide reasonable alternatives (U.S. Department of Justice & U.S. Equal Employment Opportunity Commission, 2022). The bank, therefore, declines to affect inference modules and, if it uses video at all, constrains analysis to job-relevant content with explicit human review and reason traceability. The stop rule is recorded in change control, and the contract memorializes the boundary. The stakes are not hypothetical: current litigation alleging AI-driven employment bias against a major HR technology vendor underscores the legal and reputational risks associated with tools that embed opaque or non-job-related signals (Rogelberg, 2025). In this light, a rights-first filter streamlines procurement by clarifying no-go categories upfront, and bias-audit and notice expectations, such as those in New York City’s AEDT regime, provide additional, enforceable guardrails (New York City Department of Consumer and Worker Protection, 2023).

For a fintech lender building an internal model that uses non-traditional signals to predict job success, each signal is tied back to a documented job analysis, and sensitive inferences and covert proxies are prohibited by design, honoring duty-against-unjust bias and duty-to-respect persons (Raghavan et al., 2020; European Union, 2024). The team commits to providing applicant-facing explanations, a staffed appeal channel, and pre-specified fairness metrics, including calibration within groups and error-rate parity, with continuous monitoring for drift (NIST, 2022; NIST, 2023; FinRegLab, 2023). Comparative outcome tracking assesses whether

access expansion occurs without systematic harm to protected groups and whether hires perform and stay at least as well as prior cohorts, reflecting the consequentialist cross-check on welfare (Berg, Burg, Gombović, & Puri, 2020; Das et al., 2023). When a trade-off arises between a small accuracy gain and a larger disparity, the decision rule favors the rights-preserving alternative with the least welfare loss, and the less-discriminatory-alternative analysis is archived for review. Real-world cautionary tales reinforce the need for feature-level discipline: poorly governed signals can encode historical bias at scale, as seen in the widely reported incident where an automated recruiter down-ranked résumés referencing women’s activities (Dastin, 2018).

Institutionalization determines whether programs endure beyond the initial enthusiasm of pilots. European law now classifies many employment systems as high-risk, requiring risk management, documentation, and post-market monitoring. ISO/IEC 42001 offers a certifiable management system that integrates policy, risk registers, impact assessments, supplier oversight, and human-in-the-loop boundaries. NIST’s framework provides socio-technical risk concepts and a lifecycle process that maps to familiar model-risk governance in banks. U.S. interagency guidance expects banks to extend controls to third-party vendors (European Union, 2024; ISO/IEC, 2023; NIST, 2023; Federal Banking Agencies, 2023). Read together, these instruments demonstrate that compliance and ethics are aligned. The same artifacts and routines that satisfy regulators also constitute the means by which duties to persons are honored.

Procurement and vendor management are where programs often succeed or fail. Supervisory insights and accountability research indicate that documentation, audit rights, access to testing artifacts, and notification prior to material changes are crucial to maintaining control over outsourced models. Periodic independent testing and explanation validation safeguard against degradation and supplier opacity (BIS-FSI, 2024; Raji et al., 2020). In practice, contracts

obligate vendors to provide evidence, audit histories, and maintain performance dashboards that include subgroup metrics. Change control is built into the delivery process, and the institution schedules routine independent tests to verify claims. Additionally, it maintains shadow models or backtesting harnesses to ensure accuracy. Owners and escalation paths are named in the contract, so accountability has an address.

Culture and leadership practices determine whether the duties are lived. Leaders who maintain decision logs, explain uncertainty, and invite critique foster psychological safety, enabling the surfacing of problems and learning. Teams without these habits treat documentation as a checkbox, undermining trust and reducing the likelihood of catching harmful drift in time. In hiring, leaders should be able to read subgroup performance tables, calibration plots, and explanation stability analyses. They should treat pause decisions as legitimate expressions of stewardship rather than as failures (Edmondson, 2019). The culture becomes the control when norms reinforce the use of artifacts, thresholds, and stop-rules as part of everyday work.

A brief comparison of frameworks at this point reinforces why the deontological sequence leads and the consequentialist cross-check disciplines. Both orientations endorse documentation, auditability, and monitoring, and both reject manipulative design and opaque exclusion. They diverge in the ordering of reasons. A rights-first approach prohibits coercive or discriminatory practices regardless of utility and makes explanation, notice, appeal, and job-relatedness non-negotiable constraints. Within those constraints, a welfare-oriented cross-check helps managers choose among compliant designs, verify sustained benefits, and manage externalities (NIST, 2022; Das et al., 2023; FinRegLab, 2023). In finance hiring, where non-discrimination, explanation, and appeal are codified expectations in law and local ordinance, the floor is neither optional nor flexible.

Much of the empirical literature and legal frameworks on fairness and explainability come from adjacent finance domains such as underwriting and pricing. While analogous, hiring has its own constraints and behaviors. Translation should proceed with care, and field evaluation should be staged, measured, and open to reversal (Fuster et al., 2022; Bartlett et al., 2022; FinRegLab, 2023). Metric improvements can remain sterile if downstream processes do not change. Error-parity gains at the model level do not guarantee equitable outcomes if appeals are inaccessible or if managers disregard reasons that invite reconsideration.

Future work should evaluate a rights-first program in the field with designs that can credibly attribute effects. A practical sketch is straightforward. Recruiter-level randomization or staggered rollouts assign some teams to the tool-assisted process and others to the business-as-usual approach. Primary endpoints are pre-registered for rights and outcomes. Rights endpoints include explanation adequacy scores from applicant surveys, the proportion of appeals meeting service levels, and the rate of redress granted when errors are identified. Outcome endpoints include quality-adjusted time to fill, post-hire performance indexed to role complexity, one - and two-year retention, diversity at the shortlist and offer stages, and applicant experience measures. Subgroup analyses are pre-specified with parity bands. Energy-per-inference is measured and reported. A decision rule is set in advance: when trade-offs are unavoidable, select the option that satisfies binding duties with the least welfare loss, and document the analysis and rationale of the alternatives.

A final contextual observation reinforces the case for a rights-first posture in finance hiring. Banks and fintechs intermediate public trust and operate within regimes that publicly articulate reasons and rights. Without explicit theory-led criteria, responsibilities diffuse, and outcomes depend on ad hoc choices rather than principled design. By appealing first to duties

and then disciplining them with outcome evidence, institutions can produce programs that are legible to applicants and employees, auditable by supervisors, and aligned with the sector's distinctive form of legitimacy. The style of governance that results is familiar to prudent institutions. It organizes human judgment around explicit constraints, shows its work through versioned artifacts, assigns named owners, and treats monitoring as a commitment rather than a slogan.

Ultimately, the most defensible claim is the simplest. AI-mediated hiring in finance must respect duties that precede outcomes while also demonstrating sustained improvements in outcomes over time. The research base indicates that models can enhance speed and consistency, and expand access when carefully managed. It also shows that without documentation, role clarity, fairness testing, and credible appeal, algorithms can repeat historical inequities with greater efficiency (Berg et al., 2020; Bartlett et al., 2022; Fuster et al., 2022; FinRegLab, 2023; NIST, 2022). Moral legitimacy, therefore, requires a deontic core operationalized through artifacts, thresholds, and named authorities, coupled with a consequentialist discipline that verifies benefits, mitigates harms, and corrects drift under changing conditions. Programs built on that combination are more likely to satisfy regulators, preserve trust, and realize the advantages of responsible automation in a domain where the stakes are unusually high.

In summary, moral legitimacy in AI-mediated hiring is a property of the entire socio-technical system, not just the model itself. A rights-first (deontological) floor sets non-negotiable constraints, including notice, explanation, job-relatedness, non-discrimination, accommodation, appeal, which are made real through versioned artifacts, pre-committed thresholds, named owners, and enforceable stop-rules. Within that protected space, a welfare-oriented (consequentialist) cross-check verifies that compliant designs improve outcomes and contain

externalities by tracking a standard panel of metrics and triggering revalidation or rollback when disparities widen. Institutional durability is achieved when these duties and tests are embedded across procurement and vendor contracts, model risk governance, documentation, and audit trails, as well as leader behaviors and environmental accounting for inference. The practical directive is to design the pipeline so that people, processes, and technology consistently produce truthful and fair reasons, reviewable and correctable decisions, and measurable improvements over time. Programs built on this sequence: rights, reasons, and results, are more likely to satisfy regulators, preserve trust, and deliver the efficiencies that finance demands without treating applicants as means to an end.

References

- Bank for International Settlements—Financial Stability Institute. (2024). *Managing explanations: How regulators can address AI explainability* (FSI Insights on policy implementation No. 24).
- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer Lending Discrimination in the FinTech Era. *Journal of Financial Economics*, 143(1), 30–56.
<https://doi.org/10.1016/j.jfineco.2021.08.011>
- Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *Review of Financial Studies*, 33(7), 2845–2897.
<https://doi.org/10.1093/rfs/hhz099>
- Brendel, A. B., Mirbabaie, M., & Lembcke, T. B. (2021). Ethical management of artificial intelligence. *Journal of Business Ethics*.
- Bühler, A. (2021). *Navigating ethical leadership in the age of AI: The ethical pocketknife*.
- Das, S., Stanton, R., & Sundaram, R. (2023). Algorithmic fairness. *Annual Review of Financial Economics*, 15, 455–480. <https://doi.org/10.1146/annurev-financial-110921-125930>
- Dastin, J. (2018, October 10). Amazon scrapped a secret AI recruiting tool that showed bias against women. *Reuters*. <https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG>
- Edmondson, A. C. (2019). *The fearless organization: Creating psychological safety in the workplace for learning, innovation, and growth*. Wiley.

European Commission High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*.

European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (*Artificial Intelligence Act*).

Federal Banking Agencies. (2023). *Interagency Guidance on Third-Party Relationships: Risk Management*.

FinRegLab. (2023). *Explainability and fairness in machine learning for credit underwriting: Policy analysis and industry survey*.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(4), 2049–2097. <https://doi.org/10.1111/jofi.13090>

Geburu, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.

ISO/IEC. (2023). *ISO/IEC 42001: Artificial intelligence management system*. International Organization for Standardization.

Kant, I. (1785/2011). *Groundwork of the metaphysics of morals* (M. Gregor & J. Timmermann, Eds.). Cambridge University Press.

Lau, T. (2021). *Banking on (artificial) intelligence: Navigating the realities of AI in financial services*.

Liautaud, S., & Sweetingham, L. (2022). *The power of ethics: How to make good choices in a complicated world*.

Mill, J. S. (1863). *Utilitarianism*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... Gebru, T. (2019). Model Cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM.

National Institute of Standards and Technology. (2022). *Towards a standard for identifying and managing bias in AI* (NIST Special Publication 1270).

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*.

New York City Department of Consumer and Worker Protection. (2023). *Automated Employment Decision Tools: Frequently Asked Questions*. <https://dcwp.nyc.gov>

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*.

Office of the Comptroller of the Currency. (2021/2025 update). *Comptroller's Handbook: Model risk management*.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 469–481). ACM.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Pitcan, M. (2020). Closing the AI accountability gap: Defining an end-to-end

framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). ACM. <https://doi.org/10.1145/3351095.3372873>

Rogelberg, S. (2025, July 5). Workday and Amazon face allegations of AI-driven employment bias and hiring discrimination. *Fortune*. <https://fortune.com/2025/07/05/workday-amazon-alleged-ai-employment-bias-hiring-discrimination/>

Ross, W. D. (1930/2002). *The right and the good* (P. Stratton-Lake, Ed.). Oxford University Press.

This paper utilized AI-augmented decision-making, specifically Grammarly (2025, <https://www.grammarly.com/grammar>), to refine grammar and tone. Additionally, ChatGPT (<https://openai.com/chatgpt/overview/>) was used to help compare the final draft of the paper to the prompt for coherence and completeness of the answer.

The final submission reflects the author's critical assessment and revision of any AI-generated suggestions. An ethical framework was employed to assess the benefits and risks associated with this action. All original writing is the Author's, unless cited otherwise.