Responsible AI and Organizational Change Leadership

Qualifying Exam Question 3: Dr. Risa Williams

Christopher Law

11/21/2025

Muskingum University

Department of Organizational Leadership

Responsible Artificial Intelligence use carries moral urgency and organizational complexity. This paper takes as a starting point a moral grammar that integrates deontological duties, consequentialist outcomes, and the virtues that cultivate practical wisdom, then asks how ethical intent becomes practice through organizational change. The central claim is that AI's speed, opacity, and continuous evolution strain traditional, episodic change models that presume stability, linear sequencing, and centralized control. Editorial and empirical work in Human Resources and leadership echo this strain and emphasize the need to interrogate claims of objectivity, actively manage lifecycle risks, and center fairness as both computational property and lived perception in high-stakes domains such as hiring (Boyd, Hunter, Intindola, & Serra, 2025; Diefenhardt, 2025; Jiang, Cai, & Wang, 2025).

To meet these conditions, the analysis draws on Bushe and Marshak's distinction between diagnostic and dialogic organization development and their later articulation of planned and generative change, together with recent leadership work on generative AI that elevates adaptive learning, sense-making, and ethical discernment as core capabilities (Bushe & Marshak, 2009, 2018; Tabata, Wildermuth, Bottomley, & Jenkins, 2025). The comparison of classical planned approaches with dialogic and complexity-attuned approaches shows where each helps or hinders responsible AI adoption, then evaluates how diagnostic and dialogic logics can be integrated into AI programs. The practical synthesis that follows draws on a rights-first, learning-fast approach. It puts rights, fairness, and accountability at the front of design and governance while cultivating rapid, disciplined learning cycles that adapt to shifting technical and social conditions.

Responsible AI begins with an ethical grammar that makes practice intelligible. A deontological lens clarifies duties that bind AI use in employment, including nondiscrimination, meaningful explanation, notice, accessibility, and avenues for appeal. Contemporary guidance

locates these duties within governance that prioritizes human well-being, privacy, and accountability, establishing a floor that organizations may not trade away for speed or efficiency (IEEE Global Initiative, 2017, 2019). Consequentialism adds an evidence discipline across the lifecycle. The question is not only whether a system complies with rules, but whether it improves outcomes such as time to fill, retention, quality, subgroup parity, and applicant experience while avoiding new externalities. Risk-based approaches in human resource management recommend an integrated, portfolio-level analysis that traces interdependencies among people, tasks, structure, and technology, then ties those insights to mitigation and monitoring in the field. Jiang, Cai, and Wang's integrated risk analysis adapts Leavitt's Diamond to precisely this work for generative AI in HR, offering a governance map that can evolve with the system and the organization (Jiang et al., 2025). Virtue ethics focuses on character and the cultivation of practical wisdom through routines that shape leaders and teams. In practice, this looks like artifact literacy and visible reasons for decisions, paired with a readiness to reverse course when lived effects diverge from intent. Recent leadership scholarship on generative AI emphasizes the same habits and treats ethical discernment and sense-making as daily practice rather than peripheral aspiration (Tabata et al., 2025).

The same ethical clarity explains why AI unsettles conventional assumptions about change. Episodic models presume a stable baseline, a bounded project, and linear sequencing from unfreezing to refreezing or through staged buy-in and reinforcement. AI resists each premise. Velocity pressures teams as tools and models update quickly, data distributions drift, and new prompt interfaces reconfigure how work is done on weekly cycles. Opacity obscures how inputs become outputs, which invites overconfidence in claims of objectivity and accuracy. A genealogical analysis of machine-learning assessment shows how normative design choices

can be naturalized by formatting candidates as inputs and rendering them as scored outputs (Diefenhardt, 2025). This stance also aligns with learning-organization principles that privilege short feedback loops and reflective inquiry over one-time change efforts, and with complexity perspectives that treat AI adoption as an adaptive process in non-linear, coupled systems.

Sociotechnical coupling means that outcomes emerge from the interaction among data, models, interfaces, policies, and human judgment, not solely from algorithms. Integrated risk analysis underscores these interdependencies and argues for structured identification, assessment, and control across the lifecycle rather than point solutions that target models in isolation (Jiang et al., 2025). Real-time legitimacy adds a fourth pressure. Expectations for fairness, transparency, and recourse are negotiated as stakeholders encounter systems in practice. Research on algorithmic reductionism warns against collapsing justice into a single metric, showing that procedural justice in HR decisions turns on visible job-relatedness and comprehensible explanations as much as on distributional parity (Newman, Fast, & Harmon, 2020). Empirical work in AI-enabled selection supports the point. Fairness perceptions hinge on outcomes and on whether procedures signal job-relatedness and provide intelligible reasons; they also depend on human involvement that candidates can see and trust (Yuan, Xing, & Zhao, 2025). Grounded studies of recruitment systems show that bias mitigation is not a one-time fix and that iterative retraining, monitoring, and collaboration between HR and HR software developers are required to manage drift and address new failure modes as they arise (Soleimani, Intezari, Arrowsmith, Pauleen, & Taskin, 2025).

Organization development offers a practical bridge from values to practice. Bushe and Marshak distinguish between diagnostic OD, which treats organizations as systems to be analyzed and adjusted through planned interventions, and dialogic OD, which treats

organizations as meaning-making systems in which narratives and conversations generate new realities (Bushe & Marshak, 2009). Both logics have a role in AI work. Diagnostic discipline enables bias testing, outcomes monitoring, and artifact-based governance that regulators and courts can recognize. Dialogic engagement legitimizes sense-making under uncertainty, distributed learning, and shared ownership of ethical commitments by the people who must live with the tools. Their later contrast between planned and generative change sharpens the fit. Planned change aligns with classic models and project plans. Generative change emphasizes interactions that create new possibilities and identities, an approach that often travels better in fast, complex environments where refreezing is not realistic (Bushe & Marshak, 2018). Leadership research on generative AI brings the bridge into contemporary focus by emphasizing adaptive learning, shared sense-making in the face of ambiguity, and ethical discernment as capabilities organizations must practice to steward AI responsibly at scale (Tabata et al., 2025).

Classical change models begin with a picture of the organization that can be measured and diagnosed, then move it through a planned sequence toward a defined future state. Lewin's unfreeze–change–refreeze schema and the staged methodologies that followed took for granted the usefulness of stable end states, linear sequencing, strong sponsorship, and control through plans, metrics, and training that typically arrive after design is settled. That grammar travels well when the task is to make a bounded intervention and hold the gains. It serves responsible AI by scaffolding traceability, approval gates, and governance artifacts that make duties and decisions visible and auditable.

Kotter's Eight Steps and ADKAR assume sequenced stages, stable end states, and top-down sponsorship. They help mobilize urgency, clarify sponsorship, and reinforce new behaviors, but they under-serve continuous drift and co-created meaning. Dialogic/generative

approaches, by contrast, presume evolving identities and distributed sense-making. They excel at rapid learning, legitimacy, and course correction, yet require a diagnostic spine (thresholds, audits, appeals) to meet legal duties. The hybrid uses Kotter/ADKAR strengths for mobilization and reinforcement, while dialogic routines sustain adaptation between "stages." For my dissertation trajectory, I would pursue a rights-first, learning-fast hybrid that integrates a diagnostic spine with a dialogic/generative engine.

Dialogic and generative approaches start from a different premise. Organizations are communities of conversation where narratives, symbols, and questions produce new possibilities. The work of change is to convene, widen, and focus those conversations in ways that shift how people make sense together. In that view, leadership is distributed, learning is continuous and experimental, and progress is judged by the quality of shared meaning as well as by adherence to plan. This stance supports responsible AI because it legitimizes co-design with affected groups, invites critique rather than suppressing it, and enables adaptation when monitoring reveals drift. Its risk is under-specification if it neglects the binding structure required to meet legal duties and to stop when harms emerge. The generative frame that Bushe and Marshak describe helps organizations balance these risks by distinguishing episodic, planned change from interactional processes that create new futures and identities in fast environments (Bushe & Marshak, 2018).

A hybrid path remains the most practical way to translate ethical intent into daily AI practice by integrating diagnostic OD's auditable artifacts and stop rules with dialogic/generative routines that institutionalize inquiry, dissent, and rapid adaptation. This hybrid best fits AI's velocity and uncertainty while maintaining enforceable, visible fairness and accountability. In this paper, that path takes the familiar summary rights-first, learning-fast, and renders it as an operable architecture that teams can run. The deontic floor anchors the structure by codifying

human rights and fair treatment as nonnegotiable constraints in employment, including notice, accessibility, job-relatedness, explanation, nondiscrimination, accommodation, and meaningful appeal. Consequentialist walls keep the building upright over time by specifying the outcome panels that matter for organizational success and growth, so compliant choices are tested for their effects on time-to-fill, quality, retention, subgroup parity, and applicant experience. Values-centered windows keep the interior livable by requiring explainability and transparency, so reasons are visible to the people who must rely on them and challenge them when needed, and diagnostic self-reflection functions as the building's mirrors by asking layered models to evaluate their own outputs against the original prompt and, more importantly, against observed results in the field. Dialogic inputs supply natural light and fresh air, as human deliberation revises what counts as job-relevant evidence and changes model behavior when lived experience reveals harms that metrics missed. Over the long term, layered engines learn to evaluate and contextualize diagnostic results while adapting to human interaction, so the architecture remains both sturdy and breathable.

The operating stack blends automated diagnostics with human judgment. At the base is the original diagnostic prompt that ranks a large applicant pool and surfaces a manageable shortlist, for example, stack ranking three thousand applications to identify five candidates for deeper review. Above it sits a diagnostic audit layer that compares the shortlist to the full population, flags demographic or pedigree outliers such as recurring school patterns or disproportionate subgroup selection, and asks why. If the pattern suggests proxy bias or drift, a stop gate holds the decision, records variance explanations, and routes the case for review. A third diagnostic layer closes the loop after hire by tracking how finalists initially ranked, noting whether they emerged from second-layer escalation, and assessing how each selection shifts

team composition and capability. Spanning these diagnostics is the human dialogic layer that sets and revises direction. When mission, vision, or values change, for example, to recognize nontraditional credentials or prioritize military experience, leaders update the purposes and constraints that guide the first three layers in plain language that recruiters and candidates can understand. An auditable prompt layer caps the stack by documenting thresholds, recording reasons, and preserving a transparent chain of evidence from the initial prompt to the final decision. In practice, this is what rights-first, learning-fast looks like: duties visible as artifacts, outcomes tested against lived effects, and a disciplined rhythm of diagnostic reflection and dialogic direction that keeps the system accountable to people as it learns.

To make the diagnostic spine legally defensible, treat documentation as an operating system rather than a binder. Each model and workflow should be tied to a versioned datasheet and model card that captures scope, intended use, subgroup performance, known limitations, and retraining cadence. A feature rationality catalog records job-relatedness evidence, proxy risks, and the rationale for any exclusion. Decision logs store the shortlist produced, the human rationale for the final choice, and any overrides with reasons. Threshold tables define parity bands, calibration targets, and stop rules, while an exception register tracks every breach, the corrective action taken, and the date of revalidation. Preserve data lineage by snapshotting training and evaluation sets and linking them to the exact model hash used for a hiring recommendation. These artifacts should have explicit owners, review dates, and attestation checkboxes to make accountability visible. When an appeal occurs, the organization can reconstruct the decision pathway from prompt to outcome and show that choices were constrained by pre-committed duties and evaluated against lived effects, rather than solely by design intent.

Legal defensibility, moral legitimacy, and reputational risk prevention require versioned documentation, predefined thresholds, and structured monitoring that can trigger rollback when parity is breached. Teams need explicit rationales for feature selection that demonstrate job-relatedness and guard against proxy effects; they need outcome panels that track time-to-fill, quality, retention, subgroup parity, and applicant experience; and they benefit from portfolio-level views of risk that surface interdependencies across people, tasks, structure, and technology. Without an integrated, cross-functional risk map, the adoption of generative tools in HR can amplify ethical, privacy, and trust risks through their interdependence, even when each appears manageable in isolation (Jiang et al., 2025). Dialogic capacity is equally necessary. Studies in AI-enabled selection show that people judge fairness by outcomes and by whether criteria and explanations feel job-relevant and comprehensible. A multi-stage HR-in-the-loop model captures this point by posing fairness questions at the data, feature, model, and post-assessment stages and by insisting that communication with applicants and follow-up be part of the system, not an afterthought (Yuan et al., 2025). Critical scholarship complements this operational advice by interrogating claims of objectivity and accuracy that can conceal normative choices and by examining how assessment pipelines format both candidates and observers (Barocas & Selbst, 2016; Kellogg, Valentine, & Christin, 2020; Raghavan, Barocas, Kleinberg, & Levy, 2020; Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019; Diefenhardt, 2025).

Explanation quality is a fairness lever, not a communications afterthought. Build it as a design stream with user testing. For applicants, keep explanations short, job-related, and concrete (“We looked for evidence of X because our analysis shows it predicts Y on this job”). Provide a “why not selected” variant that names the few highest-weight criteria and offers actionable next steps without revealing protected-class inferences. For recruiters and hiring managers, add

deeper “how it works” briefs, global feature importance, and standard failure modes to help them credibly challenge outputs. Run comprehension checks (readability targets, screen-reader compatibility, multi-language coverage) and short scenario quizzes to verify that audiences interpret reasons as intended. In appeals, pair the case-specific rationale with a link to the governing artifact (the model card section and the feature catalog entry) so people can see the rule behind the rationale. Treat these materials as living artifacts tied to monitoring: if parity shifts or applicant feedback raises confusion, explanations change alongside the models.

The resulting tension between diagnostic and dialogic is productive. Diagnostic logic pushes toward standardization and measurable thresholds. Dialogic logic pushes toward situated sense-making and evolving commitments. The tension prevents ethical reductionism, a familiar failure mode in which fairness collapses to a single metric and procedural justice erodes because people cannot see the job-related reasons behind the numbers. Evidence from behavioral research on algorithmic decision-making in HR supports this caution. It shows why explanations, visible criteria, and human review remain central to perceived fairness, even when statistical parity improves (Newman et al., 2020). Practitioner guidance aligns with this operating posture by urging CHROs to routinize change, reinvent HR operating models for AI, and build capability for continuous adaptation rather than episodic initiatives (Gartner, 2025).

Selection research that centers on practitioner–developer collaboration strengthens the case for this synthesis. Grounded studies show that bias mitigation is iterative and cross-disciplinary, with HR professionals and AI developers sharing accountability for data, features, models, and workflows. Perceived fairness improves when job-relatedness is made explicit and when explanations are designed for the audiences who must rely on them (Soleimani et al., 2025; Yuan et al., 2025). The hybrid’s division of labor follows: a diagnostic spine ensures systematic

mitigation and monitoring across the lifecycle through integrated, cross-functional risk governance (Jiang et al., 2025), while a dialogic engine keeps the human meaning of fair in view by convening the conversations where people interpret impacts and revise practice (Bushe & Marshak, 2018).

Translating these insights into practice begins by drawing people who live with the tools into the purpose-and-duty-setting conversation, then encoding the results in artifacts and gates that can be audited. In selection, that sequence means convening HR, hiring managers, candidate advocates, legal, data science, and DEI leads to settle why automation is being used, which rights are nonnegotiable, and what success looks like. Agreements then show up as a feature-rationality catalog that documents job-relatedness and flags likely proxies, parity bands, and stop rules set in advance for subgroup performance, as well as candidate-facing notices and explanations tested for clarity and accessibility. Reporting on a fairness panel that includes speed, quality, retention, subgroup representation, and applicant experience resists reductionism and reduces the risk of decontextualized algorithmic judgments (Newman et al., 2020). From there, responsibility travels with the lifecycle. At the data stage, diversified sourcing, careful handling of missingness and labeling, and documentation of assumptions are fairness questions rather than technical housekeeping, limiting the data humans need to review and focusing them on the "why" rather than the "how." At the feature stage, job-relatedness evidence and plain-language rationales build procedural justice. At the model stage, teams make trade-offs explicit, test in drift-reflective conditions, and plan the retraining cadence up front to match risk. At the workflow stage, interfaces that temper unreflective acceptance of scores, require reasons, and log overrides turn human-in-the-loop into a learnable practice. Empirical accounts suggest that this HR-in-the-loop

rhythm improves perceived fairness and technical quality over time (Soleimani et al., 2025; Yuan et al., 2025).

Data choices are fairness choices. Establish sourcing standards that diversify channels, document eligibility rules, and make the treatment of missingness auditable. At ingestion, tag records with cohort and time identifiers to enable drift detection. During training and evaluation, report stability metrics for both utility (e.g., precision, calibration) and equity (e.g., error parity across subgroups) over rolling windows so the team can see movement rather than one-off snapshots. In production, set a cadence for health checks that compares new shortlists to the eligible pool for pedigree skews, subgroup shifts, and unexplained variance. When a monitor breaches a preset band, the stop gate pauses automated progression, opens an entry in the exception register, and routes the case to review with a standard set of diagnostic plots and counterfactual "what if" probes. The point is not to eliminate movement but to separate benign seasonality from drifts that require retraining, feature re-justification, or interface changes.

Leadership capabilities then land in specific skills rather than abstractions. Adaptive learning and ethical discernment become visible in artifact literacy, dialogic facilitation, participatory evaluation, and risk governance. Hybrid governance lives or dies on everyday fluency. Offer a compact capability track that leaders can actually run: (1) artifact-literacy labs where participants practice reading datasheets, model cards, and parity panels and answer short applied questions; (2) decision labs that rehearse override and escalation choices using realistic dossiers and time constraints; (3) dialogic facilitation drills to convene ethics huddles, surface dissent, and translate feedback into testable changes; and (4) participatory evaluation sprints where HR and developers co-review feature choices against job analysis and candidate feedback. Each module ends with a "use at desk" job aid (question checklists, parity cheat sheets, appeal

scripts). Track capability by sampling real decisions for rationale quality and by surveying applicants and recruiters on explanation clarity and perceived recourse. Tie completion to role expectations so skills remain visible and valued rather than extra work.

The emerging leadership literature on generative AI makes that translation explicit, treating these skills as daily disciplines and tying them to trust dynamics between human and artificial agents in real work (Tabata et al., 2025; Larson & DeChurch, 2020; Cortellazzo, Bruni, & Zampieri, 2019). Two behavioral findings mark the boundary conditions. Experimental evidence indicates that employees are less likely to comply with unethical instructions from AI supervisors than from human supervisors, suggesting that organizations cannot transfer moral authority to machines and must preserve human justification and explanation for high-stakes choices (Schlager, Bhattacharjee, & Häubl, 2023). At the same time, a large-scale behavioral study shows that AI-generated advice increases dishonest behavior, underscoring the need for transparency, deliberation, and review whenever AI assists judgment in ethically charged situations (Leib, Köbis, Rilke, Hagens, & Irlenbusch, 2024). Vendor governance and documentation translate these lessons into enforceable practice through pre-committed thresholds, audit clauses, post-deployment monitoring, and incident response timelines treated as go-live gates, complemented by employer data-governance standards that prioritize agency and comprehensible explanation (IEEE Global Initiative, 2017, 2019).

Contracts should encode the rights floor and the learning cadence. Require (a) documented model scope and intended use; (b) access to model cards, evaluation datasets, and parity plots for internal audit; (c) pre-committed fairness bands and rollback criteria; (d) retraining and monitoring cadence aligned to your risk register, not the vendor release cycle; (e) incident-response SLAs that define what counts as an incident, notification timelines, and data

needed for diagnosis; (f) rights to run red-team exercises and publish high-level findings; and (g) data rights that forbid secondary use of applicant data without explicit consent. Include termination-for-cause language if the vendor blocks audit, refuses to remediate breached thresholds, or changes scope unilaterally. Treat quarterly governance as an operating meeting, not a demo: review exceptions, parity trends, appeal metrics, and upcoming changes against your people–task–structure–technology risk map so decisions reflect interdependence rather than single-feature tweaks.

Used deliberately, AI tools can strengthen dialogic learning rather than displace it. Language models can help summarize learning reviews, draft hypotheses for why a disparity widened, propose user-testing protocols for notices, and support recruiters in crafting nontechnical explanations for candidates. Leadership research suggests these gains materialize when organizations invest in AI literacy, human-centered governance, and trust-preserving routines rather than in ad hoc tool use (Tabata et al., 2025). Scenario exploration can also benefit from generative tools that simulate likely stakeholder narratives, helping prepare facilitators to include the right voices in the correct sequence without substituting simulations for real conversation. Knowledge-capture assistants that tag and retrieve datasheets, model cards, and parity analyses at decision time improve artifact literacy where it matters most.

In practice, a responsible operating model for AI-mediated selection rests on a diagnostic spine and a dialogic engine that run together. The spine registers every tool with versioning, owners, and change logs; completes datasheets and model cards that disclose scope, subgroup performance, and limitations; maintains a feature-rationality catalog grounded in job analysis; and sets parity bands and outcome panels on a fixed cadence. Publicly defensible notice, explanation, and accessibility artifacts are usability-tested with applicants, and a non-automated

path is available upon request. A risk map and response playbook, built on people–task–structure–technology interdependencies, guide escalation when thresholds are breached (Jiang et al., 2025; Newman et al., 2020). The dialogic engine convenes cross-functional ethics huddles to review exceptions and applicant feedback, runs participatory design sprints with HR, hiring managers, and candidate representatives to iterate on features and explanations, and schedules red-team probes of failure modes with visible findings.

Red teaming turns abstract risk into actionable learning. Build a shared catalog of failure modes relevant to selection: data contamination (duplicate or synthetic résumés), prompt manipulation, out-of-distribution candidates, proxy features masquerading as job-related signals, and interface designs that nudge uncritical acceptance. For each release, run targeted exercises: adversarial résumés that exploit formatting; "what if" sweeps that vary a single attribute to test invariance; subgroup stress tests at the shortlist and offer stages; and shadow deployments that compare human-only to human-plus-model pipelines. Rate findings on severity and exploitability, assign owners, and time-box fixes. Close the loop with a blameless review that updates artifacts (feature catalog, model-card limits, explanation text) and, when needed, narrows the scope until a safer variant is ready. Publishing a brief, plain-language digest of red-team findings in the transparency report helps convert internal learning into external trust.

Blameless learning reviews update thresholds, features, and artifacts, and short coaching modules build managers' fluency in artifact literacy, fairness trade-offs, and candidate conversations. A public transparency report, subject to legal review, closes the loop by sharing metrics, mitigations, and lessons. Trust grows when people can see and use recourse. Publish a concise transparency report each release cycle that covers: purpose and scope; significant changes since last release; performance and parity panels at shortlist and offer stages;

explanation and notice updates; incident summaries and mitigations; and service-level performance for appeals (acknowledgment time, resolution time, reversal rate). Establish a candidate advisory panel that meets twice yearly to review notices, explanation templates, and appeal pathways for clarity and accessibility, with participation stipends to honor time and expertise. Internally, host open "ask us anything" sessions for recruiters and hiring managers after significant changes to align on how the system works and when to escalate. Finally, provide a non-automated path upon request and report usage rates, so stakeholders can see that automation supports rather than substitutes for human judgment.

A staggered field rollout illustrates how these elements work together. Teams pre-register rights endpoints such as explanation adequacy and appeal service levels, alongside outcome endpoints for time to fill, job quality, and retention, subgroup representation at shortlist and offer stages, and applicant experience. Decision rules are set in advance. When alternatives are compatible with the duties, the option that delivers the best outcome is selected, and the reasons are recorded. When a fairness threshold is breached, rollback and a learning review are triggered rather than a quiet tweak (Jiang et al., 2025; Yuan et al., 2025).

Change theory strengthens the fit. Because AI systems evolve through nonlinear interactions across data, tools, roles, and perspectives, and through complexity, short feedback loops, experimentation, and reflection help organizations sustain ethical alignment when stable end states are unlikely. Dialogic and generative insights sustain curiosity and shared meaning as models, interfaces, and norms evolve. That learning posture is not sufficient on its own unless it is coupled with diagnostic gates that create traceability and accountability for high-stakes decisions. Bushe and Marshak's distinction between diagnostic and dialogic OD, and their articulation of planned and generative change, clarify why responsible AI needs both: the

planned discipline that produces auditable artifacts and stop rules, and the generative routines that institutionalize inquiry, dissent, and adaptation as regular work (Bushe & Marshak, 2009, 2018). Practitioner priorities point in the same direction by urging CHROs to make continuous change a core operating capability rather than a series of projects (Gartner, 2025).

The hybrid carries costs and risks: documentation and monitoring add capacity demands; metrics can induce measurement myopia; legal and regulatory shifts can render artifacts obsolete; and dialogic work requires trust and time in already stretched teams. The operating model mitigates these by using a risk-based cadence (more oversight where risk concentrates), publishing transparent reasoning that reduces defensive postures, and investing in a compact capability track so leaders can execute the habits at their desks. Red team weeks and blameless reviews counteract metric tunnel vision by surfacing unmeasured harms, while stop rules and rollback criteria prevent harm escalation when signals breach bands. These constraints keep experimentation disciplined rather than ad hoc.

An evaluative field study, rather than further concept work, should test the hybrid. Use a stepped-wedge rollout in which clusters (e.g., recruiter pods or business units) cross over from business-as-usual to the hybrid at staggered intervals, so each serves as its own control. Pre-register endpoints that mirror the operating model: rights endpoints (notice comprehension, explanation adequacy, documented job-relatedness, accessibility compliance, appeal service levels) and results endpoints (time to fill, quality/retention proxies, subgroup representation at shortlist and offer, calibration and error parity, applicant experience). Pair dashboards with mixed methods: validated measures of procedural and informational justice, recruiter/applicant trust, and artifact-literacy comprehension checks; plus qualitative coding of appeal narratives and ethics huddle notes for themes such as job-relatedness, intelligibility, and perceived recourse.

Power the study for operationally meaningful effects (e.g., a 10–15% gain in explanation comprehension or movement of subgroup error gaps into a predefined band). Share interim reads in the governance forum and reserve space in the transparency report for "what changed because of what we learned," so evidence loops back into practice.

To attribute effects, include two comparison arms, purely diagnostic and purely dialogic, with identical endpoints. In the hybrid arm, institutionalize the routines that define the model (cross-functional ethics huddles, red-team reviews, blameless learning reviews) and the audit trail (model cards, datasheets, feature rationality catalogs). Add portfolio-level risk diagnostics by adapting a people–task–structure–technology map and using Delphi-style panels to set monitoring cadences and escalation pathways that can actually run (Jiang et al., 2025). Make leadership capability development a second study strand: short, practice-proximate modules in artifact literacy, dialogic facilitation, and participatory evaluation, assessed longitudinally to test whether capability gains mediate improvements in fairness perception and performance (Soleimani et al., 2025; Yuan et al., 2025). Because legitimacy is negotiated in real time, include an ethics-of-use lens on AI assistance, examining when summarization and scenario generation help learning without eroding autonomy, privacy, or accountability, and where human review should be placed to preserve moral authority in judgment-intensive tasks (Tabata et al., 2025). Together, these elements test not only whether the hybrid outperforms single-mode programs but also why, by tracing how diagnostic structure and dialogic sense-making jointly produce trustworthy, adaptive systems.

Across the evidence and operational detail, five elements stand out. First, teams need everyday capabilities in artifact literacy, dialogic facilitation, participatory evaluation, and lifecycle risk governance. Second, structure matters: pair a diagnostic spine: datasheets, model

cards, parity bands, exception registers, and appeals, with a dialogic engine, ethics huddles, participatory sprints, red-teaming, and blameless reviews. Third, run to a cadence of pre-committed thresholds and monitoring cycles aligned to risk rather than vendor releases. Fourth, build legitimacy through explanations and notices designed and user-tested for the audiences who must rely on them, with a non-automated path available on request. Fifth, anchor governance in a cross-functional HR–AI council with authority over risk registers, rollback, and transparency reporting. Together, these elements keep rights visible, reasons examinable, and results actionable, ensuring that AI innovation remains answerable to human purpose over time. Together, these elements keep rights visible, reasons examinable, and results actionable, ensuring that AI innovation remains answerable to human purpose over time.

References

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Boyd, N., Hunter, J., Intindola, M., & Serra, T. (2025). Interdisciplinary narratives on artificial intelligence and personnel selection systems. *The International Journal of Human Resource Management, 36*(14), 2455–2479.

Bushe, G. R., & Marshak, R. J. (2009). Revisioning organization development: Diagnostic and dialogic premises and patterns of practice. *The Journal of Applied Behavioral Science, 45*(3), 348–368.

Bushe, G. R., & Marshak, R. J. (2018). Planned and generative change. *OD Practitioner, 50*(4), 9–14.

Cortellazzo, L., Bruni, E., & Zampieri, R. (2019). The role of leadership in a digitalized world: A review. *Frontiers in Psychology, 10*, 1938. https://doi.org/10.3389/fpsyg.2019.01938

Diefenhardt, F. (2025). Automating the managerial gaze: Critical and genealogical notes on machine learning in personnel assessment. *The International Journal of Human Resource Management, 36*(14), 2516–2549.

Gartner. (2025). *Top HR trends and CHRO priorities for 2026* [Webpage and webinar]. Retrieved from https://www.gartner.com/en/insights/human-resources

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (Version 2). IEEE Standards Association.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (1st ed.). IEEE Standards Association.

Jiang, Y., Cai, Z., & Wang, X. (2025). Leverage generative AI for human resource management: Integrated risk analysis approach. *The International Journal of Human Resource Management, 36*(11), 1929–1959.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Annual Review of Sociology, 46*, 365–389. https://doi.org/10.1146/annurev-soc-121919-054945

Larson, L., & DeChurch, L. A. (2020). Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *The Leadership Quarterly, 31*(1), 101377. https://doi.org/10.1016/j.leaqua.2019.101377

Leib, M., Köbis, N. C., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2024). Corrupted by algorithms? How AI-generated and human-written advice affect ethical behavior. *The Economic Journal, 134*(661), 125–154. https://doi.org/10.1093/ej/uead056

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias is not fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes, 160*, 149–167.

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). https://doi.org/10.1145/3351095.3372828

Schlager, T., Bhattacharjee, A., & Häubl, G. (2023). Employees are more likely to follow unethical instructions from humans than from AI. *Journal of Experimental Psychology: Applied*. Advance online publication.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). https://doi.org/10.1145/3287560.3287598

Soleimani, M., Intezari, A., Arrowsmith, J., Pauleen, D. J., & Taskin, N. (2025). Reducing AI bias in recruitment and selection: An integrative grounded approach. *The International Journal of Human Resource Management, 36*(14), 2480–2515.

Tabata, M., Wildermuth, C., Bottomley, K., & Jenkins, D. (2025). Generative AI integration in leadership practice: Foundations, challenges, and opportunities. *Journal of Leadership Studies, 18*(4), 41–52.

Yuan, S., Xing, L., & Zhao, D. (2025). A multi-stage HR in the loop approach to enhance fairness perceptions of AI selection systems. *The International Journal of Human Resource Management, 36*(14), 2623–2658.

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 used to assess the benefits and risks of this action. All original writing is the Author’s, unless cited otherwise.