administrative sciences

MDPI

Article

Metricized Scaffolds: AI-Enabled Simulation Training and Transfer in Human Resource Development

Nina Xie [1,*], Yujun Liu [2], Yuanyuan Wang [3] and Yiduo Wang [2]

[1] Department of Management, Faculty of Business, Lingnan University, 8 Castle Peak Road, Tuen Mun, Hong Kong
[2] Faculty of Business, Lingnan University, 8 Castle Peak Road, Tuen Mun, Hong Kong; yujunliu@ln.hk (Y.L.); yiduowang2@ln.edu.hk (Y.W.)
[3] Management and Strategy, Lee Shau Kee School of Business and Administration, Hong Kong Metropolitan University, 30 Good Shepherd Street, Ho Man Tin, Kowloon, Hong Kong; cyywang@hkmu.edu.hk
* Correspondence: ninaxie@ln.edu.hk

Abstract

AI-enabled simulation platforms are increasingly used for talent development, offering scalable, low-risk practice and algorithmic feedback for interpersonal HR capabilities such as difficult conversations, negotiation, and consultation. Yet HRD evaluation cautions that within-platform improvement may not transfer to authentic interactions once dashboards and metrics disappear. We conceptualize AI-enabled simulation training as metricized scaffolding and integrate scaffolding theory with exploration–exploitation learning dynamics to explain when algorithmic supports fade into self-regulation rather than produce metric adaptation (score chasing). We introduce Algorithmic Feedback Literacy (AFL)—learners' capability to interpret algorithmic feedback, calibrate its authority, and translate cues into portable self-scaffolds—as a mechanism linking metricized practice to transfer. We test this model in an AI-enabled virtual internship and career development program that combines repeated simulation attempts with an authentic consultation assessment. Using platform traces (592 attempts nested within 90 learners), instructor-rated consultation performance as a transfer outcome, and 102 learner reflections, we examine how distributed practice span and practice dose related to simulation gains and transfer, and how consultation complexity changes, which interactional behaviors predicted success. Distributed practice span predicted transfer beyond within-platform gains, while practice dose operated primarily through measurable simulation improvement. Complexity shifted performance relevance toward governance and boundary-setting behaviors. Qualitative evidence identified two developmental pathways: productive internalization, where learners converted metric cues into routines and mode-switching strategies, and metric adaptation, where learners optimized scores without internalizing broader professional judgment. We discuss design implications for AI-enabled HRD and retention-oriented capability building: cultivate AFL through transparency and calibration supports, design for deliberate fading, and pair metricized refinement with authentic assessments to strengthen transfer.

Keywords: human resource development; artificial intelligence; simulation-based training; transfer; algorithmic feedback literacy; metricized scaffolding; exploration–exploitation; professional judgment; talent development; retention

Check for updates

Received: 15 February 2026
Revised: 15 March 2026
Accepted: 20 March 2026
Published: 3 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In the digital age, organizations increasingly use internet communication tools and artificial intelligence (AI) to recruit, develop, and retain talent. HRD and HRD scholarship documents rapid expansion of AI- and automation-enabled development systems, including virtual HRD environments and AI-supported coaching and career guidance (Bennett, 2014; Ekuma, 2024; McWhorter, 2023; Terblanche, 2021). AI-enabled simulation platforms are one such technology: interactive systems that simulate stakeholders, deliver feedback, and scaffold repeated practice for interpersonal capabilities such as difficult conversations, negotiation, and consultation (Li et al., 2025; Wilhelm et al., 2025). These interpersonal capabilities shape how employees experience fairness, support, and voice, and thus sit upstream of retention-relevant outcomes such as well-being, job satisfaction, and trust (Aguinis & Kraiger, 2009; Hom et al., 2017). This developmental relevance also aligns with employability research, which treats capability building as a dynamic interplay between employee agency and employer supported opportunity over time (Fugate et al., 2021). At the same time, scalable AI-enabled training raises new questions about authenticity, transparency, and the limits of metricized feedback as a guide for professional judgment—especially when developmental systems also function as evaluation systems (Chen, 2024; Ekuma, 2024).

We conceptualize AI-enabled simulation training as metricized scaffolding—repeated performance practice guided by rubric-like diagnostics, scores, and rapid feedback loops. Scaffolding theory suggests that AI simulations are developmental when supports reduce ambiguity and surface criteria, then fade as responsibility shifts to the learner (Wood et al., 1976). From this perspective, AI simulation platforms offer a plausible path to scale deliberate practice for interpersonal competence—if the platforms support functions that are eventually internalized rather than permanent crutches.

Yet HRD evaluation scholarship has long warned that improvement during training—especially in tightly specified environments—does not automatically translate into transfer, defined as durable enactment of capability in more authentic settings where cues, stakes, and constraints differ from the training context (Holton, 1996; Holton & Naquin, 2005). Garavan et al. (2019) further argue that credible claims about training impact require explicit theorizing about mechanisms and boundary conditions, not only "did scores go up." These concerns are amplified in AI-mediated training because performance scores, dashboards, and AI-coded indicators continuously render performance visible and can steer attention toward what the system recognizes, not necessarily what professional judgment requires (Chen, 2024; Speer et al., 2025).

More broadly, research on algorithmic evaluation, management, and governance shows that AI in organizations increasingly spans both how humans use AI to manage and how humans are managed by AI, creating sociotechnical systems that reshape practice toward what is measurable, and correctable (Kellogg et al., 2020; Ravid et al., 2020; Pachidi et al., 2021; Hillebrand et al., 2025; Marti et al., 2024; Leavitt et al., 2025). The question, then, is whether learners internalize algorithmic cues as self-scaffolds or remain dependent on the platform. We address that question by integrating exploration–exploitation dynamics with Algorithmic Feedback Literacy (AFL).

But HR-facing interpersonal work often demands more than refinement. As ambiguity, tradeoffs, and ethical boundaries increase, competence depends on exploratory judgment and governance behavior that preserve trust and legitimacy (Herold et al., 2025; Vaccaro et al., 2024). We define Algorithmic Feedback Literacy (AFL) as the learner's capability to interpret algorithmic cues, calibrate their authority, and translate them into portable self-scaffolds and professional judgment. In this way, AFL extends organizational AI literacy by emphasizing the conceptual, practical, and ethical capabilities required to use AI

appropriately in development settings (Benlian & Pinski, 2025). This narrower, feedback-specific capability complements broader HRD calls to develop workforce AI competencies for AI-empowered environments, where employees must learn not only how to use AI tools but also how to interpret and govern them appropriately (Shi, 2025). In our framework, AFL means interpreting algorithmic cues, calibrating their authority, and translating them into portable self-scaffolds (Winstone et al., 2017; Molloy et al., 2020; Benlian & Pinski, 2025; Wood et al., 1976; Vygotsky, 1978). Without calibration and translation, metricized practice can become metric adaptation: learners optimize what the system rewards, defer rather than reflect, and risk symbolic compliance rather than robust competence (Kellogg et al., 2020; Sundaresan & Guler, 2025; Vaccaro et al., 2024).

Empirically, we study an AI-enabled virtual internship and career development program that combines repeated simulation attempts with an authentic consultation assessment. Using quantitative evidence on practice exposure, distributed practice span, task complexity, and transfer together with 102 learner reflections, we examine when repeated practice supports transfer and when it yields metric adaptation instead. In doing so, we extend HRD transfer theory by showing why platform gains are not equivalent to transfer (Holton, 1996) and speak to the design of responsible AI-enabled development as generative AI spreads across workplace training and performance settings (Brynjolfsson et al., 2025; Chen, 2024).

2. Theory and Hypotheses

2.1. AI-Enabled Developmental Infrastructures and Metricized Scaffolding

Management scholars increasingly theorize AI systems as sociotechnical arrangements that capture activity, convert it into evaluative data, and shape subsequent behavior (Kellogg et al., 2020; Pachidi et al., 2021; Zhang et al., 2025). Rather than simply automating decisions, algorithmic systems organize work by (a) capturing activity, (b) translating it into data, (c) generating evaluative signals, and (d) shaping subsequent behavior through recommendations, constraints, and incentives (Kellogg et al., 2020). This view aligns with the electronic performance monitoring (EPM) literature, which conceptualizes monitoring as technologically mediated observation of performance-relevant behavior and emphasizes that its effects depend on system features (e.g., transparency, purpose) and recipients' interpretations (Ravid et al., 2020). While much of this literature focuses on control and compliance, many contemporary AI applications are also designed for development, using repeated practice, visible metrics, and corrective feedback to support performance improvement (Ekuma, 2024; Korzynski et al., 2024; Terblanche, 2021; Chen, 2024).

HRD scholarship likewise points to the growing use of AI-enabled developmental systems—including virtual HRD environments and generative AI tools that integrate needs diagnosis, practice, and feedback—but argues that their developmental value cannot be presumed; it must be theorized in terms of underlying mechanisms, governance conditions, and transfer outcomes (McWhorter, 2023; Ekuma, 2024; Korzynski et al., 2024; Chen, 2024). Field evidence further suggests that generative AI can accelerate novice performance by making high-performing routines more available, while also raising the risk that learners orient toward what the system recommends or can readily evaluate (Brynjolfsson et al., 2025).

We conceptualize AI-enabled simulation platforms as AI-mediated developmental infrastructures: systems that repeatedly elicit performance, metricize and visualize performance through scores and indicators, and close feedback loops by providing rapid guidance for revision and reattempt (Kellogg et al., 2020; Ravid et al., 2020; Ekuma, 2024). We refer to these metric-based, iterative supports as metricized scaffolding. This framing foregrounds the paper's core tension: metricized feedback can accelerate improvement

on what is measured while displacing contextual judgment and transfer (Kellogg et al., 2020; Ravid et al., 2020). This concern is timely because LLM-enabled role-play and conversational training tools increasingly target interpersonal competencies through repeated practice and feedback, making the boundary between transferable professional judgment and platform-shaped performance particularly salient (Li et al., 2025; Wilhelm et al., 2025).

This tension is especially consequential when the target capability is interpersonal HR practice—difficult conversations, negotiation, and consultation—where quality depends on both structured competence (clarity, sequencing, action planning) and interpretive competence (sensemaking under ambiguity, ethical boundary setting, legitimacy, and fairness framing). In this context, the central theoretical problem is not whether learners' platform scores improve, but whether and how metricized scaffolding is appropriated into transferable professional judgment rather than narrow metric adaptation.

2.2. Scaffolding and Simulation-Based Learning in HRD: From Other-Regulation to Self-Scaffolding

Scaffolding theory explains how learners perform beyond current competence when supports are contingently calibrated and then faded as responsibility shifts to the learner (Wood et al., 1976; van de Pol et al., 2010). In Vygotskian terms, scaffolding operates within the zone of proximal development, enabling performance that is later internalized (or appropriated) as self-regulatory competence (Vygotsky, 1978). In AI platforms, these supports are embedded in rubrics, prompts, and metrics rather than delivered by a human tutor. This recursive structure also aligns with experiential learning theory, which conceptualizes learning as a cycle of experiencing, reflecting, thinking, and acting that becomes more effective when learners move across these modes rather than remaining in only one (Kolb & Kolb, 2005).First, scaffolding is not synonymous with generic 'support'; it is a design logic in which assistance is diagnostic and criteria-revealing (contingent), reduces ambiguity to enable deliberate practice, and is intended to be temporary rather than permanently substitutive (van de Pol et al., 2010; Pea, 2004).

Meta-analytic evidence suggests that simulation-based learning and computer-based scaffolding are most effective for complex skills when practice is structured as feedback–revision cycles and when supports strengthen strategic and metacognitive regulation rather than substitute for reasoning (Chernikova et al., 2020; Belland et al., 2017; Kim et al., 2018; Doo et al., 2020). This logic is consistent with training research that locates learning in recursive feedback loops, in which performance outcomes, revised goals, and subsequent feedback jointly drive development over time (Eller et al., 2022). Scaffolding interventions are more effective when supports are aligned with task demands and strengthen learners' strategic and metacognitive regulation rather than substituting for their reasoning (Belland et al., 2017; Kim et al., 2018; Doo et al., 2020). In AI-enabled simulations, these features appear as rubric-like diagnostics and prioritized improvement cues (contingent guidance), combined with repeated opportunities to rehearse, revise, and reattempt (conditions that enable fading and transfer of responsibility over time).

At the same time, scaffolding theory clarifies why platform learning may fail to transfer. If learners treat platform cues as compliance targets rather than criteria to be internalized, performance gains may remain scaffold-dependent and bound to what the system scores. For developmental infrastructures, the key question is whether learners convert platform cues into portable self-scaffolds—decision rules, scripts, and reflective routines that can guide action when the platform is absent. This conversion is not guaranteed by exposure to feedback; it depends on learners' capacity to interpret, calibrate, and translate algorithmic cues so that external scaffolding fades into self-scaffolding (Pea, 2004; van de Pol et al., 2010).

2.3. Transfer as the North-Star Outcome: HRD Evaluation and the Limits of Platform Learning

HRD scholarship has repeatedly warned that training evaluation can over-privilege proximal indicators (e.g., immediate improvement) while under-theorizing and under-testing the mechanisms that produce transfer—durable capability enacted in authentic contexts where cues, stakes, and constraints differ from training (Baldwin & Ford, 1988; Holton, 1996; Blume et al., 2010; Grossman & Salas, 2011). Garavan et al. (2019) further argue that credible claims about training impact require explicit theorizing about mechanisms and boundary conditions, not only whether "scores went up". These concerns are amplified in AI-mediated training and coaching because practice is continuously accompanied by platform scores, indicators, and algorithmic recommendations that can shape what learners attend to and optimize (Chen, 2024; Ekuma, 2024; Brynjolfsson et al., 2025).

AI-enabled simulations create abundant performance traces, but HRD evaluation cautions against equating "more data" with "better evidence" of transferable capability. Two validity threats are especially salient in metricized training contexts. First, platform metrics may capture only a subset of professional competence—especially when the targets include interpretive judgment, ethical reasoning, and relational legitimacy that are difficult to reduce to detectable signals. Second, learners may improve by optimizing to the platform's scoring logic without internalizing the underlying professional standards. In other words, a developmental infrastructure can produce both (a) capability development that transfers and (b) local metric optimization that remains bound to the evaluation system. A theory of AI-mediated development must therefore specify the mechanisms by which learners use algorithmic feedback to internalize standards and build self-scaffolds that persist beyond the platform—rather than merely improving what is scored. Because AI-enabled developmental infrastructures both train and score performance, the validity and governance of AI-coded indicators become conditions under which developmental claims are credible (Chen, 2024; Speer et al., 2025; Marti et al., 2024).

2.4. Exploration–Exploitation Dynamics Under Metricized Feedback

We leverage the exploration–exploitation distinction to specify the learning dynamics of metricized feedback (March, 1991). Exploitation involves refinement, efficiency, incremental improvement, and reliable execution of known routines. Exploration involves variation, experimentation, reframing, and adaptive judgment under uncertainty. March (1991) emphasizes that systems that overinvest in exploitation can achieve short-term gains while becoming brittle, narrow, and self-reinforcing. Conversely, systems that overinvest in exploration may fail to consolidate learning and may incur high costs.

Metricized developmental infrastructures tend to privilege exploitation because scores and dashboards render performance as discrete, improvable gaps. This logic is consistent with evidence that AI assistance is especially useful when it takes over codified, repetitive portions of work and leaves humans to handle higher order problem solving, although such gains may be stronger for more skilled workers than for novices (Jia et al., 2024). Dashboards, rubric-like scores, and rapid corrective cues make gaps visible and encourage iterative refinement. This is developmentally valuable in professional communication tasks where novices benefit from predictable structure—coherent openings, clear sequencing, explicit summaries, and well-formed next steps. For many learners, such exploitative refinement yields rapid improvements in clarity and procedural completeness because the platform reduces ambiguity and provides actionable cues. Related generative-AI evidence suggests that human-guided AI can improve overall solution quality while shifting the novelty–viability balance, reinforcing the idea that AI may support structured refinement more readily than open-ended exploration (Boussioux et al., 2024).

However, high-stakes HR interactions also demand exploration. Difficult conversations and consultations often involve emotionally charged circumstances, contested interpretations, and legitimacy concerns (e.g., fairness explanations, procedural justice framing, boundary setting). Negotiation and advising involve tradeoffs, incomplete information, and shifting constraints, and emerging evidence suggests generative AI tools can reshape how people prepare for and conduct negotiation-related work—intensifying the need for human judgment about strategy, legitimacy, and relational governance (Herold et al., 2025). In these contexts, high-quality performance requires sensemaking, reframing, and mode switching between solution-generation and relational governance. These exploratory competencies exceed routine refinement and are harder to capture with surface-near metrics.

This argument also aligns with research on human–AI complementarity and trust, which shows that performance depends on how interpretive authority is allocated around AI outputs (Vaccaro et al., 2024; Raisch & Fomina, 2025; Leavitt et al., 2025). Work on human–AI ensembles similarly suggests that human and AI outputs can be productively combined even when neither side alone dominates, provided the interface is designed for aggregation and judgment rather than blind substitution (Choudhary et al., 2025). Applied to AI-mediated training, the implication is that metricized infrastructures should be especially effective at accelerating exploitative refinement. Transfer under higher ambiguity, however, should depend on whether learners can engage exploration and governance behaviors beyond what platform metrics directly incentivize.

2.5. Algorithmic Feedback Literacy: A Bridge from Metricized Scaffolding to Transfer

A central question follows: what enables learners to use metricized feedback for refinement without outsourcing judgment to the system? We introduce Algorithmic Feedback Literacy (AFL) as a bridging capability that links algorithmic scaffolding to self-scaffolding and transfer by enabling learners to make algorithmic cues usable, bounded, and internalizable. AFL builds on feedback literacy frameworks that conceptualize feedback use as an agentic process involving understanding standards, making judgments, regulating engagement, and translating information into future action (Winstone et al., 2017; Molloy et al., 2020). AFL specifies what changes when the feedback source is algorithmic: learners must infer meaning from partial signals and decide how much authority to grant them. Conceptually, AFL can be viewed as a feedback-specific form of organizational AI literacy: it combines conceptual understanding of algorithmic limits with ethical boundary awareness and practical skill in using AI outputs without outsourcing professional judgment (Benlian & Pinski, 2025; Chen, 2024).

First, AFL involves interpretation: learners infer what indicators mean, how feedback maps onto professional criteria, and which behaviors likely drove the score (Kellogg et al., 2020). Second, AFL involves calibration: learners develop the capacity to decide when to trust, adapt to, or override algorithmic judgments. This calibration challenge is heightened because trust in AI is shaped not only by perceived capability but also by how agentic the system appears to be; when AI is treated more like an evaluator than a neutral tool, the need to calibrate trust becomes even more important (Vanneste & Puranam, 2025). Calibration is a key safeguard against metric adaptation: rather than treating the algorithm as an unquestionable authority, learners treat it as an informative but imperfect evaluator (Vaccaro et al., 2024). Third, AFL involves translation: learners convert metric cues into portable self-scaffolds—scripts, checklists, question stems, and reflective prompts—that can be deployed when the platform is absent. Translation enables scaffolding to fade into self-scaffolding, consistent with scaffolding theory's internalization/appropriation logic (Wood et al., 1976; Vygotsky, 1978). AFL matters because it separates productive

internalization from score chasing: metricized systems can steer attention toward what is measurable, but learners with higher AFL use algorithmic feedback to build self-scaffolds rather than outsource judgment (Ravid et al., 2020; Sundaresan & Guler, 2025).

2.6. Complexity and Professional Governance as Boundary Conditions for Transfer

Transfer performance is shaped not only by what learners have practiced but also by what the authentic task demands. As ambiguity and stakes increase, higher-quality performance requires not only exploitative actionability (clear solutions, structured next steps) but also exploration (probing, reframing, tradeoff articulation) and professional governance behaviors (boundary setting, fairness framing, ethical reasoning, appropriate referral). In March's (1991) terms, increasing complexity reallocates the value of exploration relative to exploitation.

In HR-facing interaction, governance behaviors are not peripheral; they are central to legitimacy and risk management. Explaining standards and fairness, clarifying what is negotiable, and setting boundaries are fundamental to competent HR practice. These behaviors may be less reliably inferred from surface metrics and more dependent on contextual judgment, increasing the likelihood that human complementarity becomes decisive under high complexity (Leavitt et al., 2025; Raisch & Fomina, 2025).

This contingency logic yields a sharper claim about AI-mediated developmental infrastructures. Metricized scaffolding can accelerate exploitative refinement across tasks, but transfer to complex, ambiguous consultations depends on whether learners can enact exploration and governance behaviors that preserve professional legitimacy when success criteria are not fully specified by the platform.

2.7. Hypotheses and Conceptual Model

Our framework treats AI simulation training as metricized scaffolding that can generate measurable in-platform improvement, but does not guarantee transfer unless learners internalize standards and develop self-scaffolding and adaptive judgment. This logic yields the following testable hypotheses.

H1: Distributed Practice Predicts Transfer

H1 (Distributed Practice → Transfer). *Learners who engage in more distributed practice during the AI simulation phase (i.e., practice spread across time rather than concentrated) will demonstrate higher transfer performance in the subsequent authentic consultation, after accounting for baseline performance.*

H2: Capability Gains Explain Why Practice Translates into Transfer

H2 (Mediation via Capability Gains). *The positive relationship between distributed practice and transfer performance will operate in part through capability gains during simulation—that is, distributed practice improves simulation performance, and these gains carry forward to better performance in the authentic consultation.*

H3: Algorithmic Feedback Literacy Predicts Capability Gains and Transfer

A core claim of the paper is that learners differ in how well they can use algorithmic feedback as a developmental resource rather than a score target. We expect those differences to show up first in learning gains and, potentially, in transfer.

H3a (AFL → Capability Gains). *Learners with higher Algorithmic Feedback Literacy (AFL)—the ability to interpret algorithmic feedback, judge how much authority to grant it, and translate it into self-scaffolds—will show larger capability gains during the simulation phase, even when baseline performance and practice exposure are held constant.*

H3b (AFL → Transfer). *Learners with higher Algorithmic Feedback Literacy (AFL) will demonstrate higher transfer performance in the authentic consultation, even when baseline performance and practice exposure are held constant.*

H4: Complexity Shifts What Predicts Transfer

Our learning-mode argument implies that as tasks become more ambiguous and consequential, performance depends less on routine refinement and more on adaptive judgment and professional boundary management.

H4 (Complexity as a Boundary Condition). *Task complexity in the authentic consultation will change which consultation behaviors are most strongly associated with transfer performance.*

H4a (High complexity). *In higher-complexity consultations, exploration behaviors (probing, reframing, generating alternatives, articulating tradeoffs) and governance/boundary-setting behaviors (clarifying scope, fairness standards, ethical framing, limits/referrals) will show stronger positive associations with transfer performance.*

H4b (Low complexity). *In lower-complexity consultations, exploitative/actionable behaviors (structured problem definition, summarizing, concrete next steps, action planning) will show stronger positive associations with transfer performance.*

Figure 1 integrates four elements of the argument—AI-mediated metricization, scaffolding and fading, transfer as the evaluation criterion, and exploration–exploitation dynamics under complexity—into a conceptual model of development and transfer.

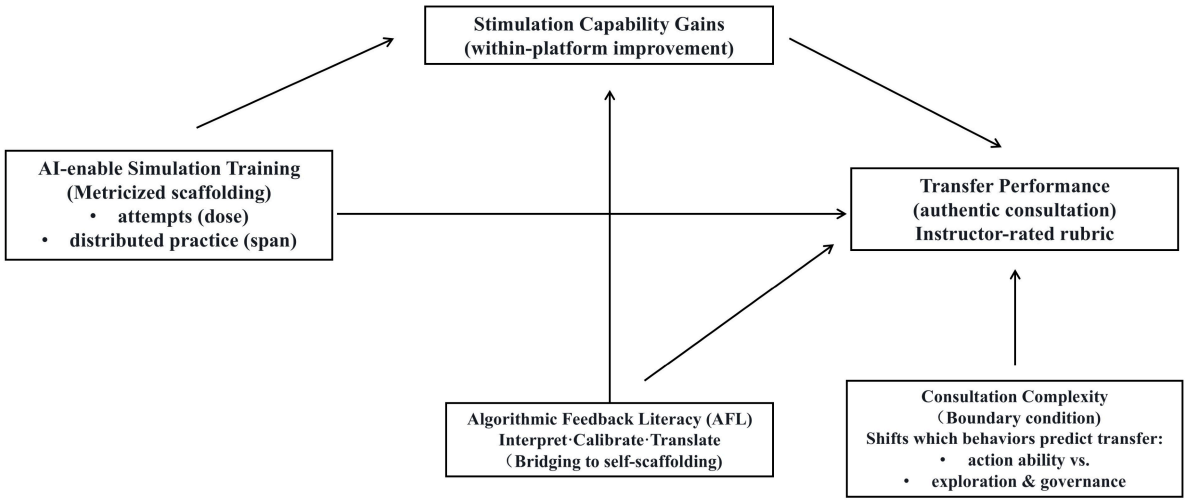


Figure 1. Theoretical model of AI-enabled developmental infrastructure and transfer.

Transfer performance is modeled as emerging in authentic consultation through in-interaction behaviors that instantiate the exploration–exploitation balance: exploitative/actionable behaviors (structured next steps, concrete solutions), exploratory behaviors (probing and reframing), and governance/boundary-setting behaviors (ethical framing, procedural fairness, and limits). Consultation task complexity moderates which behavior set most strongly predicts transfer performance, reallocating value toward exploration and governance as ambiguity increases.

3. Materials and Methods

3.1. Research Setting and Design

We study an AI-enabled virtual internship that combines repeated simulation practice with an authentic consultation assessment (Kellogg et al., 2020; Ravid et al., 2020; Ekuma, 2024; McWhorter, 2023).

The program operates as a virtual HRD environment in which learners practice HR- and career-relevant interpersonal communication through structured scenarios, rubric-aligned feedback, and iterative revision opportunities (Bennett & McWhorter, 2022; McWhorter, 2010; Chernikova et al., 2020).

We use a mixed-methods design in which quantitative analyses estimate practice–gain–transfer relationships and qualitative analyses trace how learners interpret and use algorithmic feedback. Quantitatively, we test hypothesized links among practice exposure, simulation-phase capability gains, and transfer performance, with task complexity as a boundary condition. Qualitatively, we trace how learners interpret, calibrate, and translate algorithmic feedback into portable routines (Algorithmic Feedback Literacy; AFL) to explain why metricized training produces either transferable competence or narrow metric adaptation (Winstone et al., 2017; Molloy et al., 2020).

The study has two phases:

- **Simulation phase (training):** repeated scenario-based practice with algorithmic feedback and reattempts within the AI Virtual Internship platform "*VirtualSpeech*". The design follows simulation-based learning logic (practice–feedback–revision cycles) and draws on scaffolding principles that emphasize contingent support and gradual internalization rather than permanent dependence (Chernikova et al., 2020; van de Pol et al., 2010; Doo et al., 2020).
- **Authentic consultation phase (transfer):** A live human–human consultation conducted with a standardized partner and assessed using an analytic rubric to capture performance beyond the training platform (Holton, 2005; Jonsson & Svingby, 2007).

Figure 2 summarizes the two-phase study design and the primary data streams used for mixed-evidence inference.

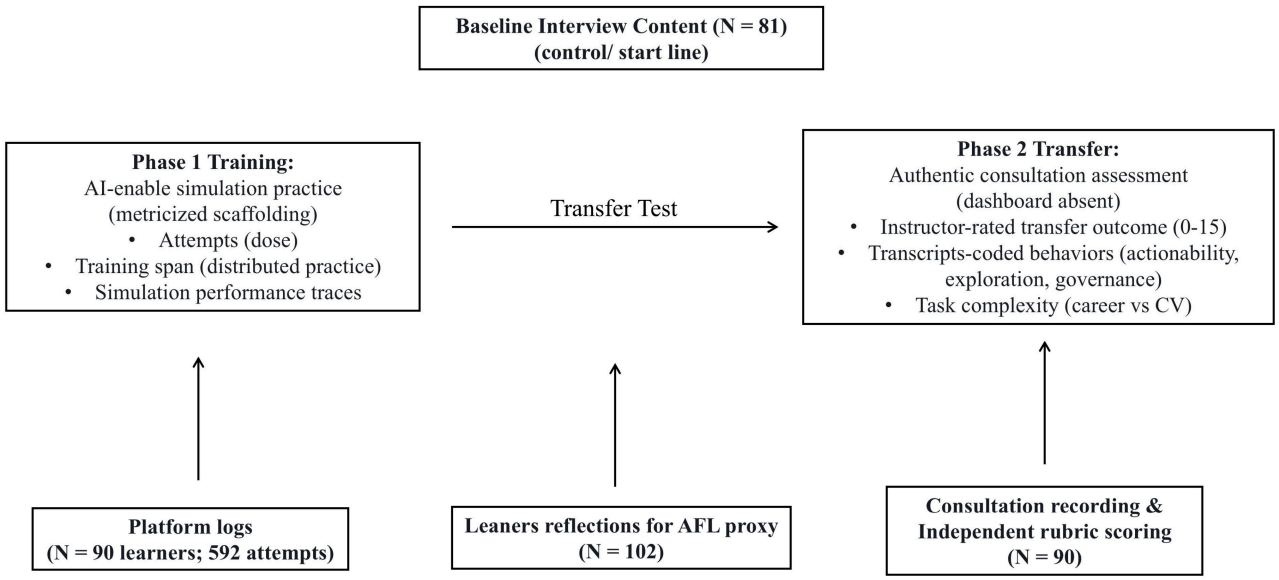


Figure 2. Study design and data sources (training phase –authentic transfer assessment).

3.2. Participants, Recruitment, and Ethics

Participants were students enrolled in a university-based virtual internship and career development program (N = 106; 81 female and 25 male; mean age = 24.15). Of these, 90 generated an analyzable platform and transferred data, and 81 also had baseline interview content available for baseline-controlled models. The setting is higher education, but the training design resembles organizational pre-entry and early-career HRD because learners practice realistic interpersonal work episodes under performance criteria and are assessed on transfer in a live consultation task. Recruitment occurred through course announcements and program mailing lists during the first week of the HRM Master Programme. Inclusion required informed consent for the research use of learning artifacts (platform logs, transcripts, reflections). Participation in the learning program was not contingent on research participation, and research scoring was separated from course evaluation to reduce perceived surveillance and reactance (Ravid et al., 2020; Chen, 2024).

We followed responsible HRD and AI research practices by (a) minimizing collection to research-relevant fields, (b) de-identifying transcripts and artifacts prior to coding, (c) restricting access to raw data to the research team, and (d) separating program evaluation from research scoring to reduce perceived monitoring threat (Ravid et al., 2020; Chen, 2024).

3.3. AI-Enabled Simulation Platform and Simulation Procedure

The simulation phase was delivered through "*VirtualSpeech*", a UK-based online training platform that combines AI-enabled conversational practice with optional virtual reality (VR) environments for professional-skills development. The platform was accessed in its web-based format from 27 August to 1 October 2025. For use with VR devices, the corresponding software version was the 2025 release. The platform supports scenario-based training for workplace interactions, including interviews, networking, difficult conversations, leadership communication, and consultation-style exchanges. Interaction was primarily conducted through AI-powered avatars that responded dynamically to participants' spoken input; for selected exercises, participants could also complete the activity in a 360-degree VR environment using a headset. As illustrated in Figure 3, learners received immediate algorithmic feedback through both a results dashboard (Panel A) and a feedback-and-recordings interface (Panel B), which together provided performance indicators, access to the audio/transcript, and reflective prompts that could be used to diagnose weaknesses and guide subsequent attempts.

Participants were assigned platform accounts and followed a structured practice pathway aligned with their designated HR role track (e.g., HR Consultant or HR Manager). Each pathway included required and recommended scenarios mapped to the course competency matrix. Before beginning a scenario, participants were encouraged to complete the associated e-learning module to review the relevant communication concepts and domain knowledge. They then launched the exercise on a computer or, where applicable, in VR, reviewed the scenario background and task instructions, selected an avatar (where relevant), and engaged in real-time spoken dialogue. After completing the conversation, participants ended the session and received immediate automated feedback.

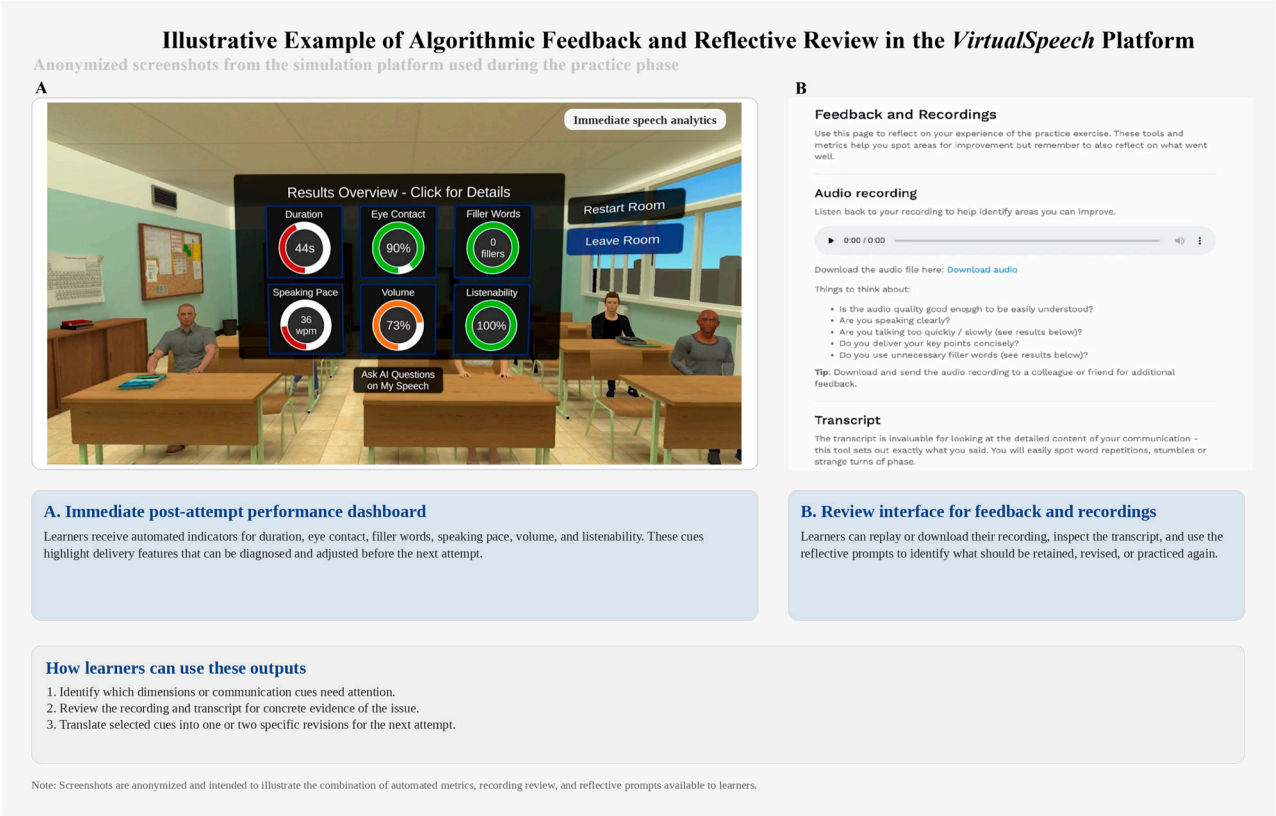


Figure 3. Illustrative Example of Algorithmic Feedback and Reflective Review in the *VirtualSpeech* platform.

3.4. Procedure

- **Baseline:** Participants completed (a) a baseline survey (demographics, prior relevant experience, AI attitudes/trust, feedback engagement) and (b) a baseline performance attempt using the same task family and rubric dimensions as later assessments, establishing a comparable starting point for gain calculations (Winstone et al., 2017; Molloy et al., 2020).
- **Simulation phase.** Learners completed multiple consultation-style scenarios on the AI Virtual Internship platform. Each scenario followed a consistent cycle: prompt → participant response → algorithmic feedback aligned to rubric dimensions → optional revision → reattempt. The platform logged scenario ID, timestamps, attempt order, and dimension-level performance indicators, enabling precise exposure measures (attempt volume and spacing) and attempt-level learning trajectories. The program included structured human-led debriefs after selected simulation activities. This emphasis on timing is consistent with coaching research showing that interventions are most effective when matched to the task cycle and the function performers most need at that moment (Hackman & Wageman, 2005). Debriefs were designed to contextualize algorithmic feedback, surface reasoning behind revisions, and encourage reflective uptake—reducing the risk of uncritical score chasing and supporting the translation of feedback into portable routines (Anseel et al., 2015; Chen, 2024; Sundaresan & Guler, 2025).
- **Transfer phase:** To assess transfer, participants completed an authentic one-to-one career and CV consultation with a tutee from another programme. After reviewing the tutee's materials and scheduling the consultation, participants conducted a recorded video session of at least 30 min focused on career direction, skills, experience, and CV improvement. Sessions were audio-recorded with auto-transcription enabled. Partic-

ipants then submitted the consultation recording and transcript, a revised CV with tracked changes, a follow-up summary email, and a brief reflection. Transfer performance was assessed using a rubric covering session compliance, needs analysis, actions and outcomes, CV revision quality, and professional communication/documentation.

3.5. Data Sources

We draw on four data streams, each tied to specific constructs in the conceptual model:

- Platform logs: scenario ID, attempt number, timestamps (submission times), revision intervals, and platform-generated dimension indicators/scores for each attempt.
- Artifacts: submitted drafts and revised versions (where revision was permitted), enabling analysis of within-scenario revision trajectories and the content of changes following feedback.
- Transcripts: full simulation interaction transcripts (prompts + learner responses + algorithmic feedback text) and authentic consultation transcripts; debrief transcripts where applicable.
- Surveys and reflections: surveys captured demographics, experience, and AI attitudes/trust; structured reflections captured how learners interpreted, calibrated, and translated algorithmic feedback (AFL) and provided process evidence distinguishing internalization from metric adaptation.

3.6. Measures

Because AI-enabled developmental infrastructures both train and score performance, the validity and governance of AI-coded indicators become conditions under which developmental claims are credible (Chen, 2024; Speer et al., 2025; Marti et al., 2024). Applied to AI-coded process indicators in developmental infrastructures, these perspectives imply construct-dependent reliability. Structured, surface-near indicators (e.g., counts, presence/absence of action steps, question frequency) should be more stable for AI coders than interpretive indicators (e.g., depth of exploration, quality of ethical boundary setting). This does not imply interpretive constructs should be abandoned; it implies they require thicker governance—human-in-the-loop auditing, rubric calibration, and transparent reporting of disagreement (Jonsson & Svingby, 2007; Wollenschläger et al., 2016; Speer et al., 2025). This governance argument is not merely methodological. It is part of the theory of developmental infrastructures: credible developmental claims depend on how algorithmic and human measurement actors are coordinated and constrained.

- **Practice exposure:** We operationalize exposure using two complementary indicators: (a) practice dose (total simulation attempts) and (b) training span (days), defined as the number of days between a learner's first and last simulation attempt. This distinguishes massed repetition from distributed engagement (Chernikova et al., 2020).
- **Capability gains (simulation phase):** We operationalized gains as Δ content, defined as the learner's mean simulation content score minus the baseline interview content score. This operationalization treats improvement as distinct from exposure volume and aligns with HRD transfer evaluation logic, which holds that proximal platform improvement is informative but not equivalent to transfer (Garavan et al., 2019; Holton, 2005).
- **Algorithmic Feedback Literacy (AFL):** AFL was operationalized as a reflection-based index coded from structured learner reflections, capturing explicit evidence of (a) interpretation, (b) calibration, and (c) translation. AFL coding data were available for 95.1% of reflections; disagreements were resolved through discussion and codebook refinement. The coded dimensions showed modest internal consistency (Cronbach's $\alpha = 0.54$). Appendix A documents coding rules and anchors. We treat this index as a

proxy (rather than a fully validated scale), which motivates cautious interpretation and future scale development.

- **Transfer performance (primary outcome):** Transfer was measured as rater-assigned rubric scores from the authentic consultation. Analytic rubrics are appropriate in this setting because they make performance criteria explicit and allow dimension-level developmental assessment rather than a single undifferentiated score (Hawk, 2009; Jonsson & Svingby, 2007). A subset of consultations was independently double-rated, and interrater reliability was assessed using an intraclass correlation coefficient (ICC). Scores were available for all 90 consultations (100%) in the analytic sample.
- **In-interaction behaviors (consultation transcripts):** Consultation transcripts from the transfer task (N = 90) were coded into actionability, exploration, and governance/boundary behavior families using a hybrid approach: AI-assisted extraction generated candidate tags at the utterance level, and human coders audited and corrected tags using a pre-specified codebook. The transcript set included 50 Chinese-language, 25 mixed-language, and 15 English-language consultations; 74 were primarily CV-focused, and 16 were broader career-focused consultations. We validated AI-assisted outputs by double-coding a subset of transcripts and quantifying agreement between human-only and hybrid-coded results, consistent with psychometric cautions that AI-based scores require explicit reliability evidence rather than assumption (Speer et al., 2025).
- **Actionability/exploitative behaviors:** structured problem definition, concrete next steps, explicit action planning.
- **Exploration behaviors:** probing, reframing, generating alternatives, articulating tradeoffs.
- **Governance/boundary behaviors:** clarifying scope/limits, fairness/standards language, ethical framing, and referral when appropriate.
- Coding is conducted with a hybrid approach (AI-assisted extraction plus targeted human checks) to ensure interpretive categories remain anchored to the coding scheme.
- **Task complexity (moderator):** Task complexity was operationalized as consultation type, coded 1 for career-direction consultations and 0 for CV-focused consultations. Complexity is treated as a boundary condition because ambiguity reallocates the value of exploration and governance relative to exploitative refinement (March, 1991).
- **Controls:** We include baseline performance, prior relevant experience, since feedback-seeking behavior is shaped not only by informational motives but also by impression-management concerns that can influence whether and how learners actively seek evaluative input. (Morrison & Bies, 1991; Anseel et al., 2015; Ravid et al., 2020).

3.7. Analysis Overview

3.7.1. Qualitative Strategy

The qualitative component explains how metricized training produces transferable capability versus metric adaptation. We analyzed reflections and selected transcripts using constant comparison logic, focusing on episodes where learners (a) interpreted feedback, (b) calibrated whether to trust it, and (c) translated it into routines that could persist beyond the platform (Molloy et al., 2020; Winstone et al., 2017). Cases were sampled for maximum variation (e.g., high vs. low AFL; high vs. low transfer; higher vs. lower complexity consultations). Coding combined theory-informed categories (AFL components; exploration vs. exploitation; feedback-seeking tactics) with emergent codes (e.g., score chasing, selective uptake, self-scaffolding routines).

Deliverables include (1) an AFL pathway typology (e.g., compliant optimizers vs. selective calibrators vs. self-scaffold builders) and (2) mechanism narratives showing how complexity shapes mode-switching, governance language, and transfer-relevant judgment.

3.7.2. Quantitative Strategy

We estimated multilevel models reflecting the nesting structure (attempts and scenarios within participants). Specifically, we modeled simulation learning using attempt-level mixed-effects growth models (592 attempts nested within 90 participants) and modeled transfer performance using participant-level regression models with pre-specified controls (N = 90). Participants completed 6.58 attempts on average (SD = 2.98; range = 1–13), spanning 6.54 unique scenarios (SD = 2.99) over an average of 35.12 days (SD = 19.20). Mediation was tested with bootstrapped indirect effects, and moderation tests used mean-centered predictors and interaction terms with simple-slope probing:

H1. *Practice exposure (volume and spacing) predicts transfer performance, controlling for baseline.*

H2. *Capability Gains mediate the practice–transfer relationship.*

H3. *AFL predicts capability gains and transfer over and above baseline and exposure.*

H4. *Task complexity moderates which in-interaction behaviors (actionability vs. exploration/governance) are most predictive of transfer.*

Analyses used the available sample for each model. Baseline-controlled models used the 81 participants with baseline data, whereas the moderation model used 90 participants. AFL missingness was represented by a missingness indicator in models that included AFL.

3.8. Mixed-Methods Integration

We integrate findings through (a) connecting, where qualitative case selection is guided by quantitative strata (AFL, gains, transfer), and (b) joint displays, linking key numeric indicators (exposure, gains, transfer, complexity, behavior codes) with qualitative process evidence. This structure strengthens construct interpretation by linking quantitative patterns (exposure, gains, transfer, complexity) with qualitative process evidence (interpretation, calibration, translation), supporting mechanism-based inference beyond proximal platform metrics.

4. Results

4.1. Qualitative Findings: A Process Model of Metricized Scaffolding-to-Transfer

We analyzed 102 learner reflections to trace how learners interpreted and used algorithmic feedback across the training cycle. The qualitative analysis is mechanism-forward: rather than cataloging themes, it clarifies how metricized scaffolding produces—or fails to produce—transferable HRD capability within an AI-mediated developmental infrastructure.

Core category: Algorithmic Feedback Literacy links metricized scaffolding to transfer.

Across reflections, Algorithmic Feedback Literacy (AFL) emerged as the bridging capability that determined whether learners turned algorithmic feedback into self-scaffolds that persisted beyond the platform. High-AFL accounts followed a consistent sequence: learners read metrics as diagnostic cues rather than verdicts, identified behavior-level causes, and converted those diagnoses into repeatable routines (e.g., scripts, checklists, practice structures). Lower-AFL accounts focused on generalized intentions (e.g., "communicate better") without specifying behavioral plans, or treat system feedback as a substitute for judgment rather than an input to it. This distinction supports our argument that metricized development produced transfer only when learners could interpret, calibrate, and translate feedback cues (Molloy et al., 2020; Winstone et al., 2017).

Figure 4 summarizes the qualitative process model and the two learner pathways.

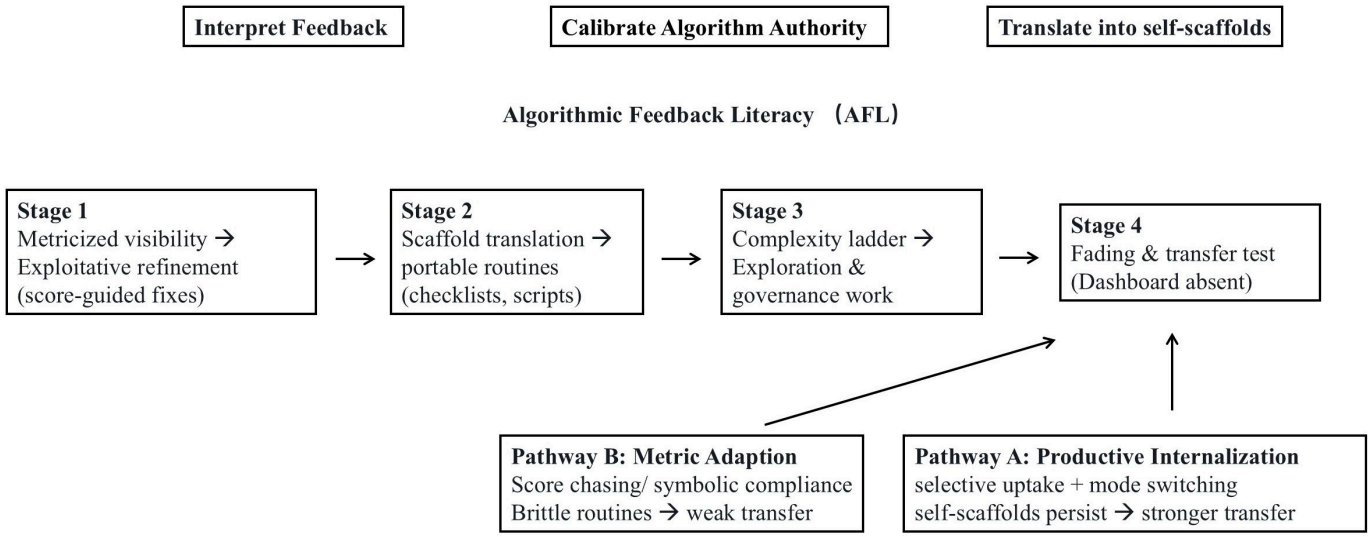


Figure 4. Qualitative process model: how metricized scaffolding produces productive internalization versus metric adaptation.

Stage 1: Metricized visibility supports exploitative refinement

Learners described dashboards and rubric-like diagnostics as making deficits legible and clarifying what to improve next. At this stage, learning was mainly exploitative: learners made direct fixes and iterative refinements aligned with platform dimensions, consistent with scaffolding that reduces ambiguity and supports revision (Wood et al., 1976; Vygotsky, 1978; Belland et al., 2017; Kim et al., 2018; Chernikova et al., 2020).

Stage 2: Scaffold translation shifts ownership toward self-scaffolding

A critical inflection point was whether learners translated platform cues into portable micro-practices. Rather than simply "trying again," learners developed routines and decision rules they could use without the dashboard. Translation was therefore the mechanism through which within-platform improvement became durable capability rather than local compliance (Holton, 1996; Holton & Naquin, 2005; Garavan et al., 2019).

Stage 3: Complexity triggers exploration and governance

As scenarios became more emotionally and ethically turbulent—conditions typical of HR practice—reflections shifted from delivery fixes to sensemaking and mode switching. Learners re-evaluated "what counts as success," emphasizing legitimacy, fairness, empathy, and boundary setting. This stage reflects exploratory learning: reframing, probing, and integrating competing constraints under ambiguity (March, 1991). Here, the limits of purely exploitative refinement became salient: clarity and structure were insufficient for high-quality performance in complex interpersonal contexts.

Stage 4: Fading reveals internalization versus metric adaptation

When dashboard salience diminished, and performance had to be held in authentic interaction, reflections diverged into two pathways. In the productive internalization pathway, learners owned their routines and self-monitored without relying on the system, indicating that scaffolding had faded into self-scaffolding (Wood et al., 1976; Vygotsky, 1978). In the metric adaptation pathway, learners prioritized score improvement and described superficial repetition, consistent with metric steering risks in monitoring and algorithmic evaluation contexts (Ravid et al., 2020) and evidence that recommendation tools can erode deep experiential consolidation when users defer to the tool (Sundaresan & Guler, 2025).

Boundary condition: Complexity reallocates transfer criteria

Across reflections, learners articulated a contingency consistent with exploration–exploitation theory: as tasks became more complex and ambiguous, exploration and governance behaviors became more decisive, while purely exploitative fixes became less diagnostic of competence (March, 1991). This grounded pattern motivated the model's complexity moderation: the same training indicators may predict transfer differently depending on whether the authentic task is structured (where actionability dominates) or ambiguous (where exploration and governance dominate).

Measurement governance

Reflections also surfaced a governance implication: learners sometimes perceive mismatches between intent and what the system detects, reinforcing that algorithmic scoring is conditional and construct-dependent (Speer et al., 2025). This aligns with an envelope perspective in which developmental infrastructures require governance arrangements that specify when algorithmic indicators can be trusted and when human oversight is needed to preserve interpretation and legitimacy (Marti et al., 2024).

4.2. Quantitative Results

4.2.1. Descriptive Statistics

Table 1 reports descriptive statistics and zero-order correlations for the focal variables. Transfer performance (instructor rubric; 0–15) averaged 12.38 (SD = 2.30), suggesting that learners generally performed relatively well on the transfer task while still showing meaningful variation. Students completed 6.58 attempts on average (SD = 2.98) across the simulation phase, with an average training span of 35.12 days (SD = 19.20), indicating substantial dispersion in simulation exposure. Transfer was positively correlated with simulation mean content performance and training span, and modestly correlated with capability gains (Δ content), suggesting that transfer was more closely aligned with developmental quality and distributed engagement than with attempt volume alone. The strong negative association between baseline interview content and capability gains also points to a room-to-improve pattern.

4.2.2. Mixed-Effects Growth Model of Simulation Performance

To operationalize simulation-phase capability development while accounting for uneven attempts and scenario differences, we estimated a mixed-effects growth model predicting attempt-level content score (0–10) from attempt order (centered at each learner's first attempt). The model included learner-level random intercepts and random slopes for attempt order, and scenario fixed effects to absorb between-scenario difficulty. The analysis used 592 attempts nested within 90 learners. Taken together, the model points to a gradual learning curve rather than a one-off jump in performance, consistent with scaffolded practice that supports incremental improvement through repeated rehearsal and feedback.

The fixed effect of attempt order was positive and statistically significant: each additional attempt was associated with a 0.085-point increase in content score (SE = 0.035, $p = 0.015$). Variance components indicated substantial between-learner differences in overall level (random-intercept variance = 0.749; residual variance = 2.080), corresponding to an intraclass correlation of approximately 0.26, while random-slope variance was close to zero, indicating limited between-learner dispersion in growth rate. Substantively, given an average of about 6–7 attempts per learner, the attempt-order effect implies modest within-person improvement attributable to repetition alone (roughly +0.4 to +0.5 points on the 0–10 content scale). Table 2 complements this pattern by showing that additional attempts predicted larger capability gains across models, whereas training span did not,

and the strongly negative baseline coefficient indicates that improvement was concentrated among learners with more room to improve; the marginal AFL coefficient in Model G2 further suggests that more reflective use of feedback may have helped some learners convert repeated practice into measurable gains.

Table 1. Descriptive statistics and zero-order correlations among focal variables.

	Mean	SD	1	2	3	4	5	6	7	8	9	10
Transfer performance (teacher total)	12.38	2.3										
Simulation content mean	5.43	1.3	0.32									
Training span (days)	35.12	19.2	0.25	0								
Total attempts	6.58	2.98	0.16	0.38	0.23							
Baseline interview content	3.36	2.05	0.07	0.45	−0.08	−0.06						
Capability gains (Δ content)	2.03	1.87	0.15	0.21	0.13	0.34	−0.78					
AFL (reflection overall)	2.89	0.68	0.1	0.2	−0.14	0.07	0.01	0.12				
Governance boundary count	1.36	1.55	0.04	0.3	−0.12	0.17	0.11	0.11	0.23			
Exploration index	2.97	0.95	0.01	0.17	−0.17	0.1	0.2	−0.09	−0.07	0.04		
Actionability index	3.72	0.77	−0.06	0.12	−0.09	0.04	0.04	0.04	−0.08	0	0.25	
Complexity (career = 1)	0.18	0.38	0.03	0.04	−0.07	0.06	0.11	−0.05	0.06	0.06	0.26	0

Note. Means and standard deviations are reported in the first two columns; zero-order correlations are shown below the diagonal. Variables are numbered in the order listed in the first column: 1 = transfer performance (teacher total), 2 = simulation content mean, 3 = training span (days), 4 = total attempts, 5 = baseline interview content, 6 = capability gains (Δ content), 7 = AFL (reflection overall), 8 = governance index, 9 = exploration index, 10 = actionability index, 11 = consultation complexity (career-direction = 1). AFL = algorithmic feedback literacy.

Table 2. Regression models predicting simulation-phase capability gains.

	G1	G2
Attempts (dose)	0.185 ***	0.171 ***
	(0.047)	(0.045)
Training span (days)	0.004	0.002
	(0.009)	(0.008)
Baseline interview content	−0.726 ***	−0.715 ***
	(0.064)	(0.067)
AFL (reflection overall)	-	0.298 †
		(0.173)
AFL missing indicator	-	−1.194
		(1.120)
Age	−0.048	−0.049
	(0.035)	(0.039)
HR experience (years)	0.112	0.093
	(0.124)	(0.119)
Gender (Male = 1)	−0.695 *	−0.615 *
	(0.317)	(0.313)
N	81	81
R^2	0.726946	0.755754

Note. Entries are unstandardized regression coefficients; standard errors are shown in parentheses. G1 includes practice exposure variables and control variables. G2 additionally includes AFL and an AFL missingness indicator. AFL = algorithmic feedback literacy. † $p < 0.10$. * $p < 0.05$. *** $p < 0.001$.

4.2.3. Scenario-Type Performance Check: Difficult Conversation (DC) Versus Negotiation (ND)

Although our core hypothesis set does not treat scenario-type differences as standalone hypotheses, we conducted a within-person check consistent with the exploration–exploitation framing (March, 1991). Among learners who completed both difficult conversation (DC) and negotiation/debating (ND) scenarios (N = 62), maximum DC performance exceeded maximum ND performance by 0.77 points on average (paired t(61) = 3.71, $p < 0.001$; paired d = 0.47). This pattern is consistent with the idea that some tasks are more amenable to metric-guided exploitative refinement than others (March, 1991), and

it complements the qualitative claim that emotionally and ethically turbulent scenarios require broader judgment beyond polished structure.

4.2.4. Hypothesis Tests

H1. *Practice Exposure Predicts Transfer.*

H1 predicted that greater distributed practice exposure during the simulation phase would be positively associated with transfer performance in the authentic consultation, controlling for baseline performance. Practice exposure was indexed by practice dose (total attempts) and distributed span (days between first and last simulation attempts). Table 3 summarizes the transfer models estimated on the baseline-available sample (N = 81), with Model T1 focusing on practice exposure, Model T2 adding capability gains (Δ content), and Model T3 adding AFL.

Table 3. Regression models predicting transfer performance in authentic consultations.

	T1	T2	T3
Baseline interview content	0.009 (0.136)	0.383 † (0.226)	0.408 † (0.242)
Attempts (dose)	0.058 (0.105)	−0.037 (0.109)	−0.041 (0.107)
Training span (days)	0.040 ** (0.015)	0.038 ** (0.014)	0.041 ** (0.015)
Capability gains (Δ content)	-	0.516 * (0.262)	0.565 * (0.275)
AFL (reflection overall)	-	-	0.093 (0.384)
AFL missing indicator	-	-	1.137 (0.991)
Age	0.112 (0.070)	0.137 * (0.065)	0.135 * (0.066)
HR experience (years)	0.066 (0.157)	0.008 (0.162)	0.002 (0.168)
Gender (Male = 1)	−1.340 † (0.716)	−0.982 (0.732)	−1.062 (0.777)
N	81	81	81
R^2	0.214424	0.261757	0.272423

Note. Entries are unstandardized regression coefficients; standard errors are shown in parentheses. T1 includes practice exposure variables and control variables. T2 additionally includes capability gains (Δ content). T3 additionally includes AFL and an AFL missingness indicator. Δ content = mean simulation content score minus baseline interview content. AFL = algorithmic feedback literacy. N = 81 for all models. † $p < 0.10$. * $p < 0.05$. ** $p < 0.01$.

Consistent with H1, training span was positively associated with transfer performance (b = 0.040, SE = 0.015, $p < 0.01$), whereas total attempts was not significant (b = 0.058, SE = 0.105, $p = 0.58$). This pattern suggests that transfer was more responsive to how practice was distributed than to how often learners simply repeated the task. In practical terms, a one-SD increase in training span (approximately 19.2 days) corresponds to about 0.77 additional points on the 15-point transfer rubric, indicating a meaningfully stronger transfer profile for learners who spread practice across time.

H2. *Capability Gains Mediate the Practice–Transfer Relationship.*

H2 predicted that the practice–transfer relationship would operate through capability gains demonstrated during simulation, consistent with HRD critiques that exposure is not itself the mechanism unless it produces portable improvement (Garavan et al., 2019; Holton, 1996; Holton & Naquin, 2005). We operationalized capability gains as Δ content,

defined as the learner's mean simulation content score minus the baseline interview content score. In other words, H2 asks whether practice matters because it generates measurable improvement rather than because exposure itself is inherently valuable.

Path a (practice → gains). In the gains model (Table 2), total attempts significantly predicted capability gains (b = 0.171, SE = 0.045, $p < 0.001$), whereas training span did not (b = 0.002, SE = 0.006, $p = 0.80$). This pattern indicates that dose, rather than span, was the main predictor of simulation gains. Baseline interview content was strongly negatively associated with gains (b = −0.715, SE = 0.067, $p < 0.001$), again suggesting a room-to-improve pattern.

Path b (gains → transfer). In the transfer model including gains (Table 3, Model 2), capability gains positively predicted transfer (b = 0.516, SE = 0.067, $p < 0.05$), controlling for baseline, dose, and span. Notably, training span remained significant (b = 0.038, SE = 0.014, $p < 0.01$), whereas attempt count remained non-significant (b = −0.037, SE = 0.109, $p = 0.70$). Thus, measurable gains carried forward into the authentic consultation, but distributed practice continued to matter even after those gains were taken into account.

Indirect effects. Bootstrapped mediation results are reported in Table 4. The estimated indirect effect of attempt count through capability gains was 0.0955 transfer points per additional attempt, with a 95% CI of [0.0019, 0.2161], supporting mediation for practice dose. In contrast, the indirect effect of training span through capability gains was negligible and not supported (95% CI included zero). The direct effect of training span remained positive (0.038; 95% CI [0.012, 0.067]). Taken together, the mediation results sharpen the distinction between dose and spacing: additional attempts mattered indirectly by raising simulation-phase capability gains, whereas spacing contributed to transfer through processes not fully captured by Δ content, such as consolidation or self-regulation.

Table 4. Bootstrapped indirect and direct effects of practice exposure on transfer performance via capability gains.

Effect	Estimate	95% Bootstrap CI (LL)	95% Bootstrap CI (UL)
Indirect (Attempts → Gains → Transfer)	0.095506	0.001938	0.216101
Indirect (Span → Gains → Transfer)	0.002093	−0.00689	0.011878
Direct (Attempts → Transfer)	−0.03702	−0.24465	0.149748
Direct (Span → Transfer)	0.038093	0.012277	0.067379

Note. Point estimates are bootstrapped for both direct and indirect effects. The final two columns report the lower and upper bounds of the 95% bootstrap confidence interval (CI). Effects whose 95% CIs do not include zero are interpreted as statistically supported. CI = confidence interval.

H3. *Algorithmic Feedback Literacy Predicts Capability Gains and Transfer.*

H3a and H3b predicted that higher Algorithmic Feedback Literacy (AFL) would be associated with greater capability gains and higher transfer performance (Molloy et al., 2020; Winstone et al., 2017). We operationalized AFL as a reflection-based index consistent with the paper's definition of AFL as calibration-and-translation capability under metricized feedback (Kellogg et al., 2020; Ravid et al., 2020). More specifically, AFL was intended to capture learners' capability to interpret, calibrate, and translate algorithmic feedback into portable self-scaffolds.

In the gains model (Table 2, Model 2), AFL showed a marginal positive association with capability gains (b = 0.298, SE = 0.173, $p < 0.10$). However, in the transfer model including AFL (Table 3, Model 3), AFL was not significantly related to transfer performance (b = 0.093, SE = 0.384, $p = 0.81$), net of baseline, dose, span, and gains. Thus, H3a received weak (marginal) support, while H3b was not supported in the current quantitative operationalization. Substantively, this pattern suggests that the reflection-based AFL

proxy captures upstream learning processes more clearly than downstream instructor-rated consultation quality.

H4. *Consultation Complexity Moderates Which Behaviors Predict Transfer.*

H4 predicted that task complexity reallocates which behavioral processes best predict transfer: In higher complexity consultations, exploration and governance/boundary-setting behaviors should be more diagnostic. In lower complexity consultations, exploitative/actionable solution behaviors should be more diagnostic (March, 1991). Using the consultation topic proxy (career-direction vs. CV; higher vs. lower complexity) and AI-coded consultation process indicators (actionability, exploration, governance boundary-setting), we estimated moderation models with interaction terms (Table 5; N = 90), controlling for simulation mean content performance, training span, attempts, AFL, and covariates. In reader terms, the key question is whether different task contexts make different behaviors more performance-relevant.

The interaction between consultation complexity and governance boundary-setting was positive and statistically significant (b = 1.410, SE = 0.682, $p < 0.05$). Simple slopes (Table 6) indicated that governance boundary-setting was positively associated with transfer under higher complexity (b = 1.268, SE = 0.646, $p = 0.053$) but not under lower complexity (b = −0.142, SE = 0.185, $p = 0.446$). In contrast, the complex interactions for actionability and exploration were not statistically significant. Taken together, Tables 5 and 6 indicate a selective rather than general moderation pattern: ambiguity elevates the value of governance and boundary management more reliably than it elevates broad actionability or exploration.

Figure 5 visualizes the simple-slopes pattern reported in Table 6 for the complexity × governance interaction.

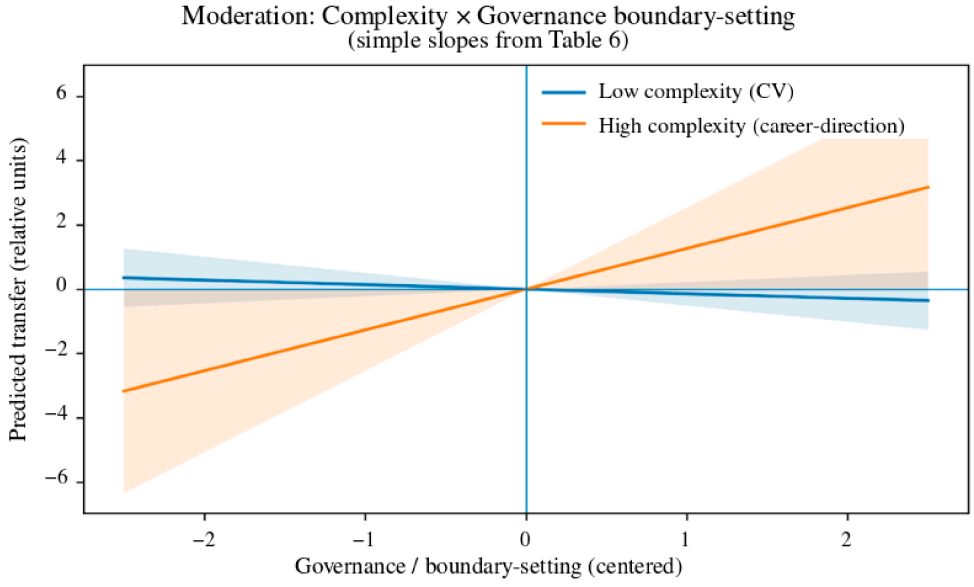


Figure 5. Complexity moderates the relationship between governance boundary-setting and transfer performance (simple slopes).

Table 5. Moderation of the associations between consultation behaviors and transfer performance by consultation complexity.

	H4_Full
Simulation content mean	0.607 **
	(0.226)
Training span (days)	0.029 *
	(0.012)
Attempts (dose)	0.002
	(0.114)
AFL (reflection overall)	0.295
	(0.365)
AFL missing indicator	0.704
	(1.195)
Complexity (career = 1)	0.343
	(0.707)
Actionability (centered)	−0.191
	(0.355)
Exploration (centered)	−0.071
	(0.370)
Governance boundary count (centered)	−0.142
	(0.185)
Complexity × Actionability	−1.148
	(1.299)
Complexity × Exploration	−0.704
	(0.833)
Complexity × Governance	1.410 *
	(0.682)
N	90
R^2	0.213349

Note. Entries are unstandardized regression coefficients; standard errors are shown in parentheses. Complexity is coded 1 = career-direction consultation and 0 = CV consultation. Actionability, exploration, and governance boundary-setting are mean-centered before the interaction terms are estimated. AFL = algorithmic feedback literacy. * $p < 0.05$. ** $p < 0.01$.

Table 6. Conditional simple slopes of consultation behaviors by consultation complexity.

Effect	Slope	SE	t	*p*
Governance (low complexity)	−0.14168	0.184927	−0.76615	0.445958
Governance (high complexity)	1.268187	0.646256	1.962359	0.053381
Exploration (low complexity)	−0.07109	0.370208	−0.19204	0.848225
Exploration (high complexity)	−0.77468	0.780189	−0.99294	0.323888
Actionability (low complexity)	−0.19114	0.355299	−0.53797	0.592172
Actionability (high complexity)	−1.33912	1.228657	−1.0899	0.279199

Note. Reported coefficients are conditional simple slopes derived from the moderation model in Table 5. Low complexity refers to CV consultations, and high complexity refers to career-direction consultations. SE = standard error.

5. Discussion and Implications

5.1. What We Learn About AI-Enabled HRD and Transfer

This study shows that AI-enabled simulations function as developmental infrastructure only when metricized feedback is internalized rather than merely followed. Quantitatively, both practice volume (attempts) and practice distribution (training span) matter, but in different ways. Attempt volume predicts simulation-phase capability gains, which, in turn, predict transfer performance in an authentic consultation. By contrast, distributed practice predicts transfer even when controlling for gains, consistent with the idea that spacing supports consolidation and self-regulation beyond within-platform score improve-

ment. At the same time, task complexity changes, which consultation behaviors are most diagnostic of transfer performance: governance and boundary-setting behaviors become more consequential under higher complexity, highlighting that professional judgment in HR-facing interactions is not reducible to "more actionable steps" alone.

5.2. Theoretical Implications

The findings extend digital-age HRD theory in three ways: by distinguishing platform gains from transfer, by showing how complexity reallocates the value of exploitative versus exploratory learning, and by identifying AFL as the mechanism that links algorithmic feedback to self-scaffolding. This mechanism-first interpretation is also consistent with the coaching literature, which shows that developmental interventions work through client characteristics, relationship processes, and contextual conditions rather than through feedback provision alone (Feldman & Lankau, 2005).

For HRD leaders and educators, AI-enabled simulations are best used as a formative layer within a broader capability pipeline, with authentic assessments reserved for judging transfer. Table 7 translates this mechanism logic into design and governance guidance: the same features that support feedback and scale can also invite over-trust, score chasing, and misplaced authority if organizations do not clearly delimit what the system captures, what it misses, and when human override is required.

Table 7. Design principles for AI-enabled simulation training in HRD (recommended practices and practices to avoid).

Design Objective	Recommended Practices (What to Do)	Practices to Avoid (What Not to Do)
Clarify the purpose of AI in training	Frame the platform as formative practice; separate developmental feedback from summative evaluation; provide learners control over when scores are visible.	Using simulation scores as selection/promotion signals; punitive monitoring; leaderboards that convert practice into competition.
Support transparency and calibration	Explain what feedback signals capture (and miss); show exemplars; encourage comparison between algorithmic feedback and human coaching where feasible.	Opaque scoring; "black box" feedback with no explanation; implying that the algorithm is fully authoritative for interpersonal judgment.
Cultivate Algorithmic Feedback Literacy (AFL)	Use reflection prompts that ask learners to interpret, calibrate, and translate feedback into routines; teach when to accept, adapt, or disregard cues.	Assuming learners will automatically know how to use feedback; leaving interpretation to guesswork; discouraging critical engagement with feedback.
Design for deliberate fading and transfer	Reduce reliance on dashboards over time; require self-generated checklists; include delayed feedback cycles; schedule spaced practice across weeks.	Persistent real-time metrics throughout all practice; constant score salience; massed practice crammed into short windows.
Balance exploitation with exploration	Mix structured drills with open-ended cases; ladder scenario complexity; reward reframing, probing, and tradeoff articulation in addition to action planning.	Single "best script" training; overly narrow rubrics that reward compliance rather than adaptive judgment; optimizing for one metric at the expense of broader capability.
Assess transfer in authentic interaction	Include at least one human–human transfer assessment with a rubric-based evaluation; give feedback on governance/boundary setting under ambiguity.	Evaluating effectiveness only through in-platform metrics; claiming transfer without an authentic performance task.

Table 7. *Cont.*

Design Objective	Recommended Practices (What to Do)	Practices to Avoid (What Not to Do)
Anticipate ethical dilemmas and escalation boundaries	Include cases involving fairness, discrimination, confidentiality, mental-health risk, harassment, and referral decisions; teach when human escalation or override is required.	Treating ethically sensitive judgments as fully automatable; using AI as the sole authority in cases with employee consequences; omitting referral pathways in high-risk situations.
Protect privacy and legitimacy	Minimize data collection, de-identify transcripts, and communicate data use; implement human oversight and audit for bias and drift; document governance "envelopes".	Collecting unnecessary personal data; deploying AI feedback without oversight; using data in ways not covered by consent; ignoring bias and fairness concerns.

Note. The table translates the study's findings into design guidance for AI-enabled HRD and highlights that effective simulation systems require both instructional design and governance safeguards.

5.3. Practices to Use and Practices to Avoid in AI-Enabled Training

Table 7 also shows that the design of AI-enabled simulations is not only an instructional issue but also a governance issue. This governance emphasis is consistent with recent HRD work framing AI implementation as a problem of nested complexity spanning technological, ethical, regulatory, and human-development layers rather than a purely technical design choice (Yawson & Goryunova, 2025). Features that support learning—such as feedback, visible scores, and scalable coding—can also encourage over-trust, score chasing, and misplaced authority if learners are not taught what the system captures and what it cannot adjudicate. From a stakeholder perspective, this is important because intelligent machines rarely create value for only one party; the same system can simultaneously benefit organizations while imposing privacy, bias, or legitimacy costs on learners and other affected stakeholders (Matthews et al., 2025). Effective programs, therefore, need clear boundaries between formative guidance and high-stakes judgment, plus escalation pathways for cases involving fairness, confidentiality, discrimination, mental-health risk, or referral. These safeguards matter not only for learning quality but also because AI systems can generate encoded, evaluative, and relational inequalities as they are developed, implemented, and used in organizations (Karunakaran et al., 2025). Framed this way, transparency and governance strengthen the authenticity and legitimacy of AI-enabled development rather than constraining it. This also resonates with management-education research showing that developmental programs are more powerful when learners can connect the process to agency, self-understanding, and authentic action rather than mere performance compliance (Byrne et al., 2023). This matters because developmental conversations are experienced through power relations as well as feedback content; unequal authority can shape whether coaching is perceived as supportive and how participants feel toward the organization (Ben-Hador, 2024). Transparency is not only an ethical principle; it is also a learning condition that enables calibration. When learners understand how feedback is produced, which signals are partial, and when human override is required, they are better positioned to use AI as a developmental aid rather than as an authoritative judge.

6. Limitations

Participants are HRM students with limited full-time workplace exposure and therefore represent pre-entry or early-career HRD rather than fully socialized role incumbents. Their consultation performance may reflect early-stage professional identity formation and emerging interactional routines rather than stable role competence, affecting both the level and variability of transfer outcomes. Limited experience may also increase

reliance on scripts, rubrics, and platform signals, potentially amplifying exploitative refinement and increasing susceptibility to algorithmic authority effects (e.g., treating system feedback as definitive rather than diagnostic). Future studies should test whether AFL operates similarly among working professionals and whether experience moderates each AFL component (interpretation, calibration, translation)—especially whether experienced workers are better able to override platform signals when contextual judgment or ethical boundary-setting is required. In addition, participation in a virtual internship may attract learners who are already more motivated or comfortable with technology-enabled learning, which could upwardly bias observed engagement and limit generalizability to mandatory workplace training.

The training scenarios and transfer assessment are anchored in HR/career-development conversations, where legitimacy, fairness explanations, and boundary-setting are central to perceived competence. Transfer mechanisms may differ in contexts with stronger incentives, accountability, or power asymmetries (e.g., sales, healthcare, leadership performance management), especially in settings where algorithmic systems are used for high-stakes evaluation rather than development. In such contexts, metric adaptation pressures may be stronger, and governance "envelopes" around how algorithmic outputs are interpreted may become decisive (Marti et al., 2024; Ravid et al., 2020). Accordingly, the present findings should be interpreted as evidence about developmental AI simulations, not as an endorsement of applying the same metricized feedback systems to selection, discipline, or performance management without additional safeguards.

Exploration behaviors, governance/boundary-setting, and AFL include interpretive components and are therefore more vulnerable to measurement error than surface-near indicators (e.g., presence of action steps). This is particularly important in AI-coded or AI-assisted measurement contexts where reliability cannot be assumed and must be demonstrated (Speer et al., 2025). Although we triangulate reflections with performance traces and rubric-based outcomes, construct validity can be strengthened through clearer behavioral anchors, expanded coder calibration procedures, and additional indicators beyond language (e.g., behavioral timing, turn-taking patterns, and decision-quality markers). Future work should also test whether coding validity differs by task complexity, since ambiguity may change the observable signatures of exploration and governance. In addition, our AFL measure is a reflection-based proxy. Reflections capture articulated sensemaking but may under-represent tacit calibration or in-the-moment judgment during live interaction, motivating future validation against behavioral and longitudinal outcomes. Transfer is assessed after the training phase but not tracked longitudinally. We therefore cannot assess whether gains persist in real workplaces or how opportunity to apply the skill, supervisory support, and transfer climate sustain or erode capability over time.

7. Future Research Directions

Future research should test four unresolved questions: whether AFL is trainable, whether complexity and experience change its value, whether transfer persists in workplace settings, and which design guardrails reduce metric adaptation without slowing learning. First, test boundary conditions by examining whether work experience, ambiguity, emotional load, and ethical stakes change the balance between exploitative refinement and exploration judgment in producing transfer. Second, test whether AFL can be taught through calibration modules, structured reflection, and deliberate fading schedules, and whether such interventions improve transfer. Third, examine longitudinal transfer and workplace enactment, following participants into internships or early employment to assess persistence and the role of transfer climate, managerial coaching, and organizational governance of AI feedback. Fourth, design and evaluate guardrails against metric adaptation—

such as delayed score visibility, narrative-first feedback, and no-metric rounds—to test how platforms can preserve learning efficiency without turning metrics into the target.

8. Conclusions

AI-enabled simulations can produce rapid within-platform improvement, but the value of these gains depends on whether metricized supports are internalized and remain usable when platform cues disappear. By conceptualizing AI simulation training as metricized scaffolding and integrating exploration–exploitation learning dynamics, this study explains why exploitative refinement is often necessary yet insufficient for authentic performance under ambiguity. Consistent with this logic, distributed practice span predicted transfer beyond within-platform gains, while practice dose operated primarily through measurable simulation improvement. As consultation complexity increased, performance relevance shifted toward governance and boundary-setting behaviors, indicating that transfer under ambiguity depends not only on polished structure but also on legitimacy-preserving judgment in interaction.

Algorithmic Feedback Literacy (AFL) clarifies the mechanism behind these patterns. AFL captures how learners interpret algorithmic signals, calibrate their authority, and translate cues into portable self-scaffolds—enabling scaffolding to fade into self-regulation rather than remaining an external target. Qualitative evidence revealed two developmental pathways: productive internalization, in which learners convert metric cues into routines and mode-switching strategies, and metric adaptation, in which learners optimize scores without internalizing broader professional judgment. Practically, effective AI-enabled HRD should be designed as a transparent developmental infrastructure rather than an opaque evaluation device: cultivate AFL through calibration supports and structured reflection, design for deliberate fading, and pair metricized refinement with authentic assessments that demand exploration and governance under realistic constraints.

Author Contributions: Conceptualization, N.X. and Y.W. (Yuanyuan Wang); methodology, N.X.; software, Y.W. (Yiduo Wang); validation, data curation, and formal analysis, N.X., Y.L. and Y.W. (Yiduo Wang); writing—original draft preparation, N.X.; writing—review and editing, N.X., and Y.L.; visualization, Y.L.; supervision and project administration, N.X.; funding acquisition, Y.W. (Yuanyuan Wang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the University Grants Committee (UGC), Hong Kong, Fund for Innovative Technology-in-Education (FITE), project title "Bridging in Web 3.0: Engaging Business Student in the Metaverse Office for Virtual Internship".

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Office of Research and Knowledge Transfer (protocol code EC136-2324 and date of approval 21 March 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. Due to ethical and privacy considerations, the original AI training data contain personally identifiable information and cannot be made publicly available. To support transparency and reproducibility, the authors will provide, upon reasonable request, the cleaned and anonymized aggregated data used for the reported analyses, together with data dictionaries and analysis code.

Acknowledgments: The authors thank *VirtualSpeech* for collaboration as the AI training platform, including guidance on virtual training set up and continuous IT support for participants. During manuscript preparation, the authors used generative AI tools to support analytic checking and editing reviews. All AI-assisted outputs were reviewed and edited by the authors, who take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviation

The following abbreviations are used in this manuscript:

AFL	Algorithmic Feedback Literacy
AI	Artificial Intelligence
CI	Confidence Interval
CV	Curriculum Vitae
DC	Difficult Conversation
EPM	Electronic Performance Monitoring
HR	Human Resources
HRD	Human Resources Development
ICC	Intraclass Correlation Coefficient
LLM	Large Language Model
ND	Negotiation/ Debating
VR	Virtual Reality

Appendix A. Algorithmic Feedback Literacy (AFL) Coding Rubric (Illustrative)

This appendix summarizes the rubric used to code structured learner reflections for evidence of Algorithmic Feedback Literacy (AFL). AFL captures how learners (a) interpret algorithmic feedback, (b) calibrate its authority, and (c) translate cues into portable self-scaffolds. Coders assessed each reflection using the anchors below and resolved disagreements through discussion. The goal is to make the qualitative-to-quantitative proxy transparent; future research can refine and validate a dedicated AFL scale.

Table A1. AFL coding rubric and illustrative anchors.

AFL Dimension	No/Low Evidence	Moderate Evidence	Strong Evidence
Interpretation (What does feedback mean?)	Repeats a score/label without linking it to behavior; treats feedback as generic or confusing.	Links at least one feedback cue to a specific behavior (e.g., "I interrupted too early"); partial understanding of criteria.	Explains how multiple cues map onto consultation goals; can articulate what would count as improvement and why.
Calibration (How much to trust it?)	Accepts feedback uncritically or rejects it wholesale; no awareness of algorithm limits or context mismatch.	Shows selective trust (e.g., accepts some cues, questions others) but limited rationale.	Explicitly articulates boundary conditions (what the system can/cannot detect); uses human judgment to weigh feedback against context.
Translation (How to use it next time?)	Expresses intent to "do better" without a plan; focuses on score improvement only.	Derives a concrete tactic or checklist item for next attempt (e.g., "ask two clarifying questions first").	Builds a reusable routine or mode-switching strategy that would apply beyond the platform (e.g., a consultation opening script plus decision rules for when to explore vs. converge).

References

Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology*, *60*, 451–474. [CrossRef] [PubMed]

Anseel, F., Beatty, A. S., Shen, W., Lievens, F., & Sackett, P. R. (2015). How are we doing after 30 years? A meta-analytic review of the antecedents and outcomes of feedback-seeking behavior. *Journal of Management*, *41*(1), 318–348. [CrossRef]

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, *41*(1), 63–105. [CrossRef]

Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. (2017). Synthesizing results from empirical research on computer-based scaffolding in STEM education: A meta-analysis. *Review of Educational Research*, *87*(2), 309–344. [CrossRef] [PubMed]

Ben-Hador, B. (2024). The practice of manager as coach (MAC): Unequal power relations and their effect on feelings toward the organization. *Human Resource Development Quarterly*, *35*(1), 67–88. [CrossRef]

Benlian, A., & Pinski, M. (2025). The AI literacy development canvas: Assessing and building AI literacy in organizations. *Business Horizons*. *in press*.

Bennett, E. E. (2014). Introducing new perspectives on virtual human resource development. *Advances in Developing Human Resources*, *16*(3), 263–280. [CrossRef]

Bennett, E. E., & McWhorter, R. R. (2022). Dancing in the paradox: Virtual human resource development, online teaching, and learning. *Advances in Developing Human Resources*, *24*(2), 99–116. [CrossRef]

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*(4), 1065–1105. [CrossRef]

Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V., & Lakhani, K. R. (2024). The crowdless future? Generative AI and creative problem-solving. *Organization Science*, *35*(5), 1589–1607. [CrossRef]

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, *140*(2), 889–942. [CrossRef]

Byrne, J., Shantz, A., & Bullough, A. (2023). What about us? Fostering authenticity in entrepreneurship education. *Academy of Management Learning & Education*, *22*(1), 4–31. [CrossRef]

Chen, Z. (2024). Responsible AI in organizational training: Applications, implications, and recommendations for future development. *Human Resource Development Review*, *23*(4), 498–521. [CrossRef]

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research*, *90*(4), 499–541. [CrossRef]

Choudhary, V., Marchetti, A., Shrestha, Y. R., & Puranam, P. (2025). Human–AI ensembles: When can they work? *Journal of Management*, *51*(2), 536–569. [CrossRef]

Doo, M. Y., Bonk, C. J., & Heo, H. (2020). A meta-analysis of scaffolding effects in online learning in higher education. *The International Review of Research in Open and Distributed Learning*, *21*(3), 60–80. [CrossRef]

Ekuma, K. (2024). Artificial intelligence and automation in human resource development: A systematic review. *Human Resource Development Review*, *23*(2), 199–229. [CrossRef]

Eller, F. J., Gielnik, M. M., Yeves, J., Alvarado, Y. C., & Guerrero, O. A. (2022). Adjusting the sails: Investigating the feedback loop of the opportunity development process in entrepreneurship training. *Academy of Management Learning & Education*, *21*(2), 209–235. [CrossRef]

Feldman, D. C., & Lankau, M. J. (2005). Executive coaching: A review and agenda for future research. *Journal of Management*, *31*(6), 829–848. [CrossRef]

Fugate, M., van der Heijden, B., De Vos, A., Forrier, A., & De Cuyper, N. (2021). Is what's past prologue? A review and agenda for contemporary employability research. *Academy of Management Annals*, *15*(1), 266–298. [CrossRef]

Garavan, T., McCarthy, A., Sheehan, M., Lai, Y., Saunders, M. N. K., Clarke, N., Carbery, R., & Shanahan, V. (2019). Measuring the organizational impact of training: The need for greater methodological rigor. *Human Resource Development Quarterly*, *30*(3), 291–309. [CrossRef]

Grossman, R., & Salas, E. (2011). The transfer of training: What really matters. *International Journal of Training and Development*, *15*(2), 103–120. [CrossRef]

Hackman, J. R., & Wageman, R. (2005). A theory of team coaching. *Academy of Management Review*, *30*(2), 269–287. [CrossRef]

Hawk, T. F. (2009). Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning. *Academy of Management Learning & Education*, *8*(4), 612–614. [CrossRef]

Herold, S., Heller, J., Rozemeijer, F., & Mahr, D. (2025). Brave new procurement deals: An experimental study of how generative artificial intelligence reshapes buyer–supplier negotiations. *Journal of Purchasing and Supply Management*, *31*, 101012. [CrossRef]

Hillebrand, L., Raisch, S., & Schad, J. (2025). Managing with artificial intelligence: An integrative framework. *Academy of Management Annals*, *19*(1), 343–375. [CrossRef]

Holton, E. F. (1996). The flawed four-level evaluation model. *Human Resource Development Quarterly*, *7*(1), 5–21. [CrossRef]

Holton, E. F. (2005). Holton's evaluation model: New evidence and construct elaborations. *Advances in Developing Human Resources*, *7*(1), 37–54. [CrossRef]

Holton, E. F., & Naquin, S. S. (2005). A critical analysis of HRD evaluation models from a decision-making perspective. *Human Resource Development Quarterly*, *16*(2), 257–280. [CrossRef]

Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology*, *102*(3), 530–545. [CrossRef] [PubMed]

Jia, N., Luo, X., Fang, Z., & Liao, C. (2024). When and how artificial intelligence augments employee creativity. *Academy of Management Journal*, *67*(1), 5–32. [CrossRef]

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. [CrossRef]

Karunakaran, A., Lebovitz, S., Narayanan, D., & Rahman, H. A. (2025). Artificial intelligence at work: An integrative perspective on the impact of AI on workplace inequality. *Academy of Management Annals*, *19*(2), 693–735. [CrossRef]

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, *14*(1), 366–410. [CrossRef]

Kim, N. J., Belland, B. R., & Walker, A. E. (2018). Effectiveness of computer-based scaffolding in the context of problem-based learning for STEM education: Bayesian meta-analysis. *Educational Psychology Review*, *30*(2), 397–429. [CrossRef]

Kolb, A. Y., & Kolb, D. A. (2005). Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of Management Learning & Education*, *4*(2), 193–212. [CrossRef]

Korzynski, P., Kim, S., & Egan, T. (2024). Bridging human resource development processes through generative artificial intelligence. *Human Resource Development Quarterly*, *35*(3), 247–256. [CrossRef]

Leavitt, K., Barnes, C. M., & Shapiro, D. L. (2025). The role of human managers within algorithmic performance management systems: A process model of employee trust in managers through reflexivity. *Academy of Management Review*, *50*(4), 745–767. [CrossRef]

Li, Z., Babar, P. P., & Peiris, R. L. (2025). Generative role-play communication training in virtual reality for autistic individuals: A study on job coach experiences in vocational training programs. In *CHI 2025 proceedings*. Association for Computing Machinery.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, *2*(1), 71–87. [CrossRef]

Marti, E., Lawrence, T. B., & Steele, C. W. (2024). Constructing envelopes: Taming disruptive algorithms. *Academy of Management Journal*, *67*(5), 1273–1301. [CrossRef]

Matthews, M. J., Su, R., Yonish, L., McClean, S., Koopman, J., & Yam, K. C. (2025). A review of artificial intelligence, algorithms, and robots through the lens of stakeholder theory. *Journal of Management*, *51*(6), 2627–2676. [CrossRef]

McWhorter, R. R. (2010). Exploring the emergence of virtual human resource development. *Advances in Developing Human Resources*, *12*(6), 623–631. [CrossRef]

McWhorter, R. R. (2023). Virtual human resource development: Definitions, challenges, and opportunities. *Human Resource Development Review*, *22*(4), 582–601. [CrossRef]

Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education*, *45*(4), 527–540. [CrossRef]

Morrison, E. W., & Bies, R. J. (1991). Impression management in the feedback-seeking process: A literature review and research agenda. *Academy of Management Review*, *16*(3), 522–541. [CrossRef]

Pachidi, S., Berends, H., Faraj, S., & Huysman, M. (2021). Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science*, *32*(1), 18–41. [CrossRef]

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences*, *13*(3), 423–451. [CrossRef]

Raisch, S., & Fomina, K. (2025). Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Academy of Management Review*, *50*(2), 441–464. [CrossRef]

Ravid, D. M., Tomczak, D. L., White, J. C., & Behrend, T. S. (2020). EPM 20/20: A review, framework, and research agenda for electronic performance monitoring. *Journal of Management*, *46*(1), 100–126. [CrossRef]

Shi, L. (2025). Global perspectives on AI competence development: Analyzing national AI strategies in education and workforce policies. *Human Resource Development Review*, *24*(4), 447–476. [CrossRef]

Speer, A. B., Oswald, F. L., & Putka, D. J. (2025). Reliability evidence for AI-based scores in organizational contexts: Applying lessons learned from psychometrics. *Organizational Research Methods*. OnlineFirst. [CrossRef]

Sundaresan, S., & Guler, I. (2025). Algorithmic recommendation tools and experiential learning in clinical care. *Organization Science*, *36*(5), 1786–1802. [CrossRef]

Terblanche, N. (2021). Coaching techniques for sustained individual change during career transitions. *Human Resource Development Quarterly*, *32*(1), 11–33. [CrossRef]

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*, 2293–2303. [CrossRef] [PubMed]

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, *22*, 271–296. [CrossRef]

Vanneste, B. S., & Puranam, P. (2025). Artificial intelligence, trust, and perceptions of agency. *Academy of Management Review*, *50*(4), 726–744. [CrossRef]

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wilhelm, L. T., Ding, X., Knutsen, K. M., Carik, B., & Rho, E. H. (2025). How managers perceive AI-assisted conversational training for workplace communication. In *CUI 2025 Proceedings*. Association for Computing Machinery.

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, *52*(1), 17–37. [CrossRef]

Wollenschläger, M., Hattie, J., Machts, N., Möller, J., & Harms, U. (2016). What makes rubrics effective in formative assessment? Transparency of performance criteria and feedback use. *Contemporary Educational Psychology*, *44–45*, 1–11. [CrossRef]

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. [CrossRef] [PubMed]

Yawson, R. M., & Goryunova, E. (2025). Nested complexity: A conceptual framework for leveraging AI for sustainable organizations and human resource development. *Advances in Developing Human Resources*, *27*(2–3), 91–123. [CrossRef]

Zhang, M. M., Cooke, F. L., Ahlstrom, D., & McNeil, N. (2025). The rise of algorithmic management and implications for work and organisations. *New Technology, Work and Employment*, *40*, 659–671. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.