



The Ghostwriter in the Machine

Undisclosed AI, the Trust Penalty, and the Future of Performance Feedback



Based on the 2026 research by Jonathan H. Westover, PhD | Nexus Institute for Work and AI

Her contribution to the project was **relentless**, demonstrating a **sharp** intellect and an **empathetic** approach to team dynamics. She consistently provided balanced feedback and appropriate solutions, ensuring that resources were optimized for maximum efficiency. The final delivery was

The Illusion of Fluency

She cannot point to anything false. She simply cannot find her manager in it.

Generative AI has quietly moved into the most personal task in management. The resulting reviews are polished, balanced, and oddly frictionless—but they substitute machine fluency for accountable managerial judgment.

This small moment of doubt is no longer an anomaly; it is a structural feature of working life.

A Structural Asymmetry in Feedback

The Manager's Reality

78% used AI for **feedback** or reviews in the prior year

73% are offered **practice** for high-stakes situations

44% enter employee names/details into public AI tools

The Asymmetry Funnel

The contribution to the performance and reputation of the individual for a meaningful and consistent look. By incorporating relevant data in the assessment for that person's role, and it has shown that there is a clear advantage to using AI tools.

The Employee's Reality

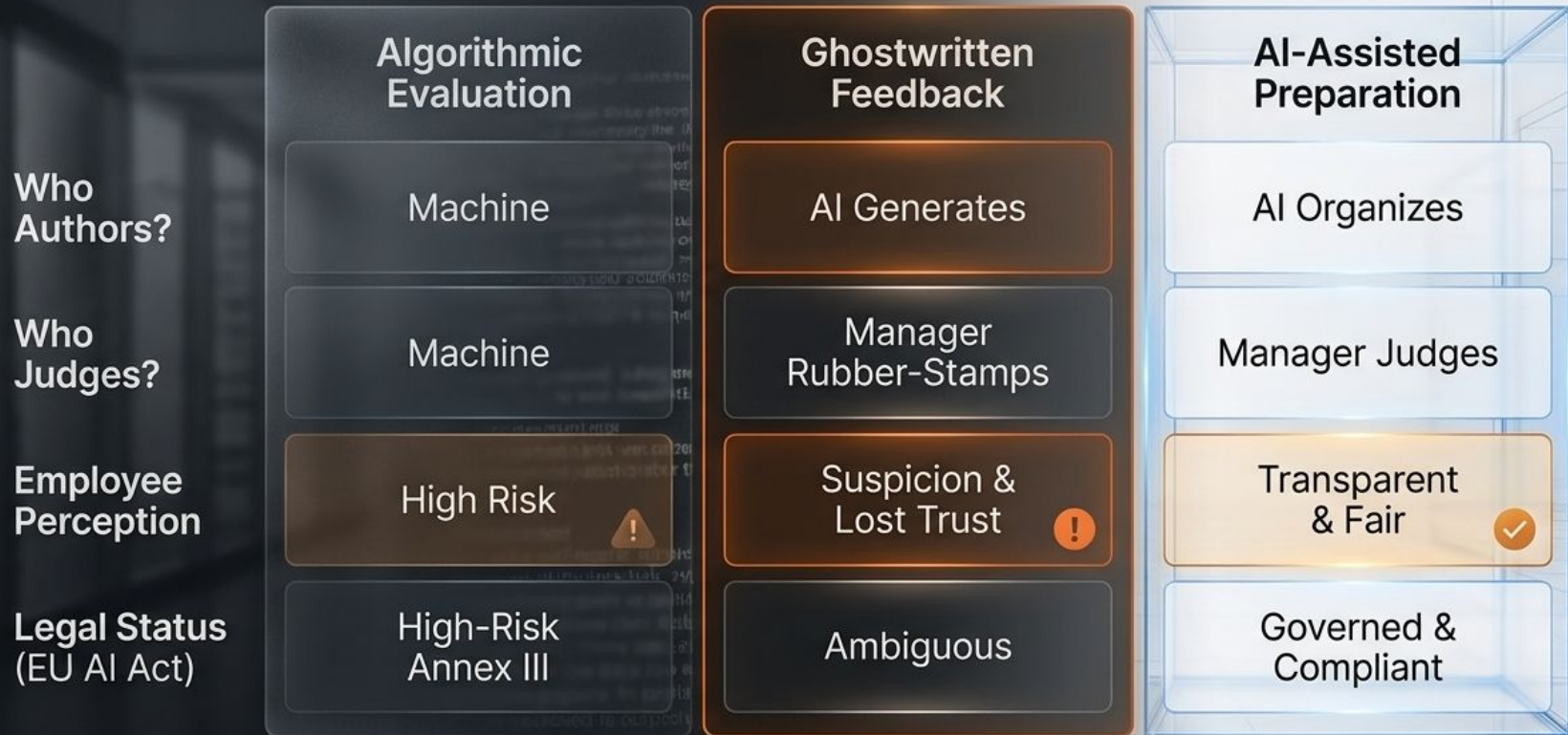
16% told AI played a role in their review

Therefore, using AI is guaranteed to improve performance and not management, and not necessarily, the relevant data feedback or the quality of the review, as the AI is not necessarily the best in being not clearly.

40% suspect undisclosed use or are unsure

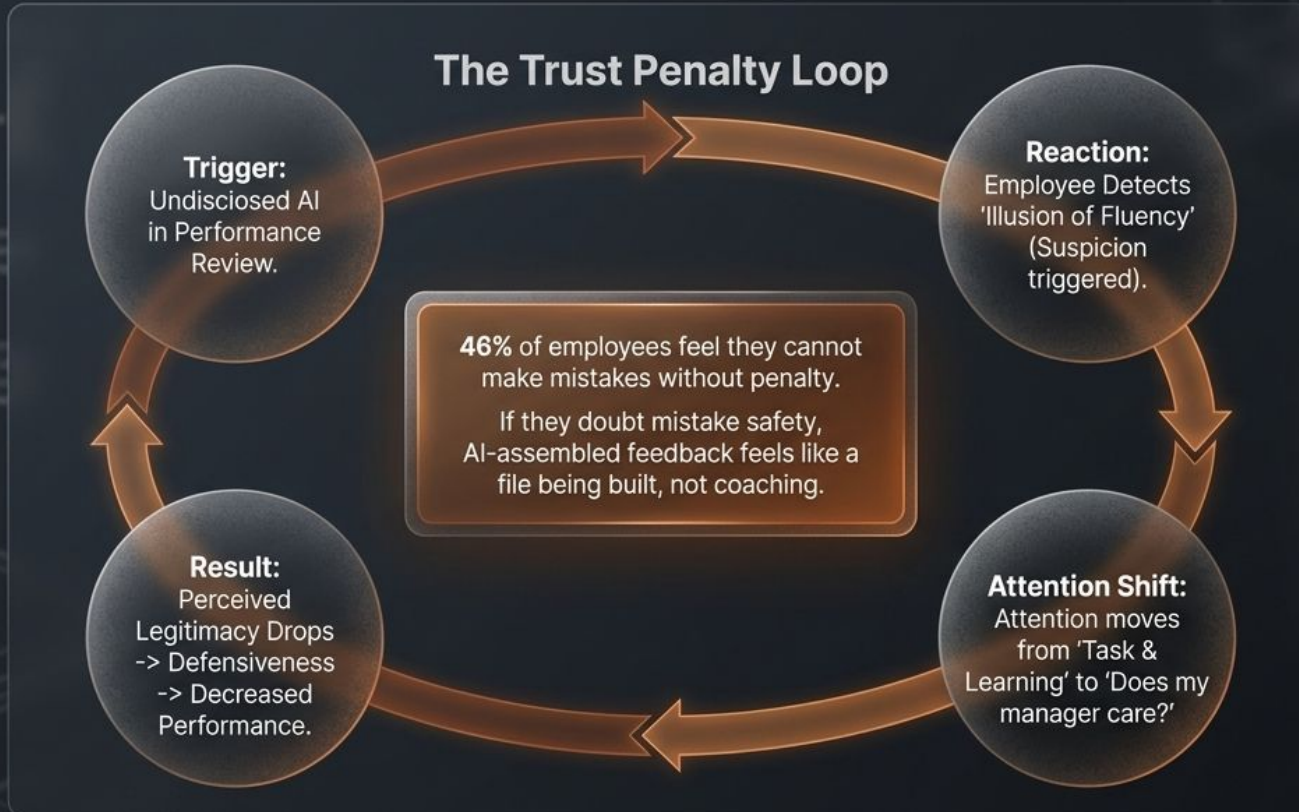
Managers have the tools and the practice. Employees receive the output with zero visibility into how it was produced. **This asymmetry** is the engine of the trust problem.

The AI Feedback Spectrum



Example: Deliveroo's rider-ranking system.

The Psychological Cost of Concealment



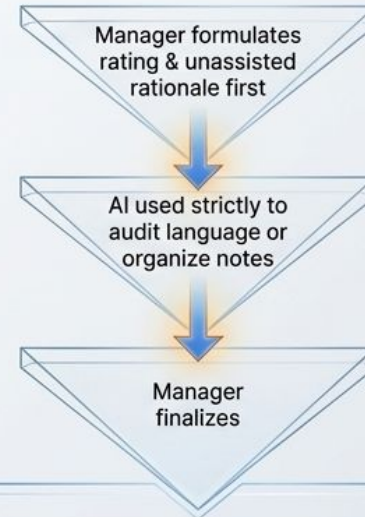
The Automation Bias Trap

Path A: The Trap



Result: Omission Errors (missed nuances) & Commission Errors (accepting AI hallucinations).

Path B: The Guardrail



Result: True Accountability & Informational Justice.

Tool-induced sameness flattens real differences between employees, making team calibration impossible.

The Bias Multiplier: AI and the Fairness Deficit

The Baseline:

Women are historically less likely to receive specific feedback tied to business outcomes, receiving vague praise or criticism instead.

Employee consistently meets expectations. However, **needs to improve communication** in team meetings and **be more assertive** in presenting ideas. **Contributions are valued**, but a **more proactive approach** would be beneficial for future leadership roles. Good job overall.

The AI Amplification:

Language models reproduce this pattern. AI frequently infers gender from job titles and adds critical length to women's feedback.



Generative models do not correct inequitable feedback patterns—they quietly scale them.

The Closing Window for Unmanaged AI



Reallocating AI: From Ghostwriter to Sparring Partner

The Human vs. AI Allocation Map

Performance Review Lifecycle



Phase 1 — Performance Review Lifecycle —>

Evidence Gathering

Uses: Summarizing year-round notes, organizing peer input.

Preparation & Rehearsal

Uses: Role-playing difficult conversations, stress-testing evidence.

Evaluative Judgment

Uses: Setting ratings, determining pay/promotion.

Delivery

Manager owns the final word and delivers accountable feedback.

The Blueprint: Disclosure & Accountability

1. Disclose by Default (The Norm, Not a Confession)

- **Why:** Exposure by a third party damages trust more than self-disclosure.
- **Action:** Mandate a standard, plain-language disclosure line across the organization. **Specify AI's exact role (e.g., organized notes) and boundaries (did not set ratings).**

2. Keep Evaluative Judgment Human & Named

- **Why:** Procedural justice requires employees to know who is responsible for judgments.
- **Action:** Managers must record ratings and unassisted rationales before opening drafting tools. **Tools must be configured so they cannot propose ratings.**

The Blueprint: Rehearsal, Audit, & Governance



Rehearse, Don't Substitute

85% of workers say rehearsal improves performance, but only 32% get to practice. Use AI to role-play the hard parts, not script the final text.



Calibrate and Audit for Bias

Use AI as a checker, not a writer. Run human-written drafts through language audits to flag vague praise or gendered personality labels.



Govern AI in People Decisions

Map where AI touches performance. Provide employees a clear appeal route for human review. Do not tie performance metrics to AI tool usage.

Operationalizing AI in Performance Management

Pitfalls

Burying AI use in a generic legal footnote.

Approving an AI-suggested rating with minor edits (Rubber-stamping).

Reading AI's generated script during the review.

Managers using public ChatGPT for reviews (**44% currently do**).

Best Practices

Stating exactly what AI did (and didn't do) in plain language.

Forcing the manager to write a 3-bullet rationale before AI is engaged.

Role-playing defensiveness with AI, then using personal notes.

Providing internal, ring-fenced enterprise tools for personnel data.

The Ultimate Goal: Reinvesting Time into Human Capability

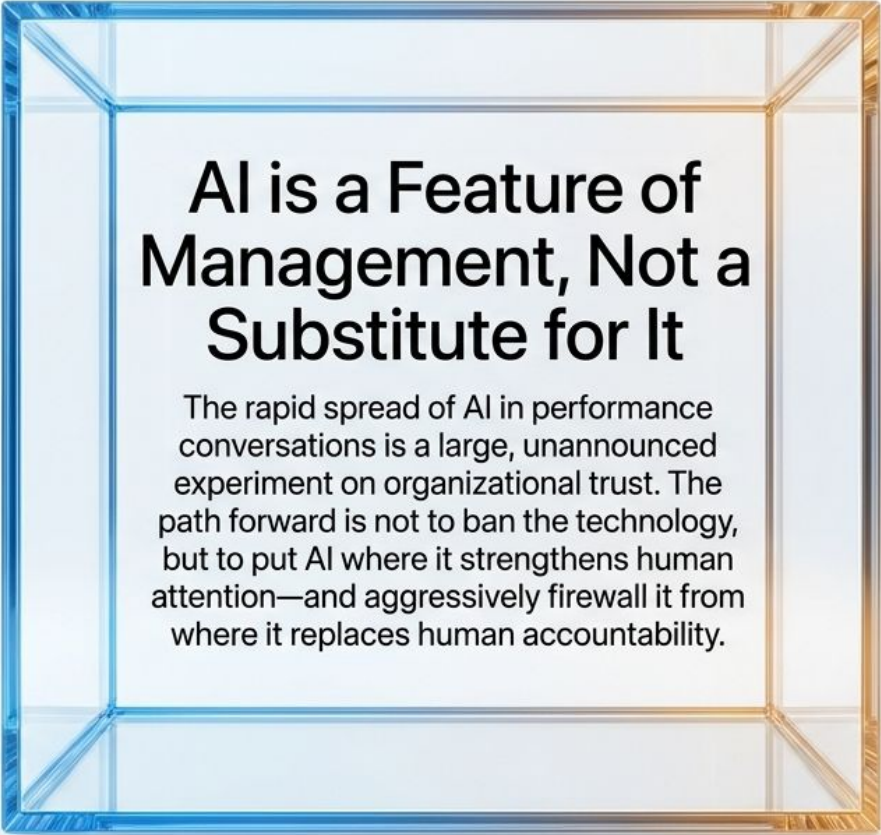
AI compresses administrative burden
(Saving 3 hours on drafting)

Expands Human Coaching Capacity
(Reinvesting time into continuous conversations)

Coaching as a Core Competency:
Assessed with the same rigor as financial skill.

Transparent Governance:
Explicit boundaries and routes to human review.

A Culture of Safe Feedback:
Measuring feedback quality from the recipient's side.



AI is a Feature of Management, Not a Substitute for It

The rapid spread of AI in performance conversations is a large, unannounced experiment on organizational trust. The path forward is not to ban the technology, but to put AI where it strengthens human attention—and aggressively firewall it from where it replaces human accountability.

A performance system must develop people, not merely document them.