Uniglobal of Journal Social Sciences and Humanities
Journal Homepage: www.ujssh.com

Beyond AI Adoption: Employee Risk Voice, Managerial Error Disclosure and Algorithmic Monitoring in Generative-AI Workflows: A Critical Integrative Review

Dai, Qian[1*], Perema Kumari[1] & Mohammad Nor Idham[1]

[1]SEGi University, Kota Damansara, 47810 Petaling Jaya, Selangor, Malaysia

*Corresponding author email: 1447043244@qq.com

To link to this article: https://doi.org/10.53797/ujssh.v5i2.20.2026

Abstract: Organisations adopting generative artificial intelligence (GenAI) require more than employee acceptance: they need workers to recognise, question and escalate erroneous, biased, privacy-sensitive or unsafe outputs. Yet research on AI at work frequently examines productivity, technology attitudes or suggestions for adoption rather than warnings about particular outputs. This critical integrative review connects research on GenAI risks and human–AI reliance with employee prohibitive voice, managerial error disclosure, psychological safety and algorithmic management. A bounded, source-verified search and purposive synthesis of peer-reviewed studies, research reviews and selected guidance documents were conducted through 8 October 2026. The literature supports the relevance of voice opportunity, interpersonal risk and manager response. Field evidence links leaders' disclosure of their own errors to employees' error sharing. Evidence directly linking managerial admission of GenAI-output errors to subsequently observed AI-risk voice remains limited. We develop a five-stage corrective-voice process (risk exposure, detection, appraisal, reporting and organisational response) and identify conditional roles for error-disclosure signals, psychological safety and punitive monitoring. The review distinguishes empirical findings from proposed links and outlines opportunity-sensitive measures and designs capable of testing rival explanations. Its principal contribution is an operationally specific account of speaking up about machine-generated risks, separate from endorsing AI implementation.

Keywords: generative artificial intelligence, prohibitive voice, psychological safety, error management, algorithmic monitoring

1. Introduction

Generative artificial intelligence (GenAI) now contributes to drafting, customer interaction, knowledge retrieval and other service work. Evidence of task-level productivity gains has accelerated organisational interest. In an experiment on professional writing tasks, access to ChatGPT led to faster completion and higher evaluated quality (Noy & Zhang, 2023). A large customer-support deployment found that gains from AI assistance differed across workers (Brynjolfsson et al., 2025). Such evidence establishes the relevance of the technology, not its infallibility. Language generation can produce plausible but ungrounded material, and risk analyses describe misinformation, discriminatory content and disclosure hazards, among other concerns (Ji et al., 2023; Weidinger et al., 2022). The organisational challenge is therefore how errors in AI-assisted work become visible before they reach decisions or services, and not only how AI tools are introduced.

Formal technical controls are indispensable; they are not always sufficient. Employees can notice a wrong recommendation, recognise a privacy implication or observe that an AI-generated response would be unsuitable for a client. The resulting warning is a discretionary communication act shaped by opportunity, competence, anticipated impact and interpersonal consequences. An employee who questions a machine-generated recommendation may simultaneously appear to question the manager who selected the tool, the work process that produced it or colleagues who relied on it. Research on employee voice and silence thus supplies a necessary social complement to work on AI reliability and human oversight (Detert & Burris, 2007; Morrison, 2014, 2023).

The closest evidence cautions against treating all AI-related voice as the same outcome. In a survey of 500 US workers, Dong et al. (2024) linked organisational listening during GenAI training to autonomy, competence, adoption

*Corresponding author: 1447043244@qq.com
https://ujssh.com/ All right reserved.

attitudes and suggestions about AI implementation. Lin and He (2024) examined AI awareness and voice in hospitality; their theoretical pathways concern the employee's relationship with AI awareness and organisational conditions, not documented reports of defective outputs. These are important studies, but encouraging AI adoption and warning against a risky AI output can involve different objects, motives and costs. A constructive warning may even oppose the immediate preferences of a manager strongly committed to GenAI deployment.

Recent studies have made nearby mechanisms more explicit. Tian et al. (2025) reported associations among leader prohibitive voice, psychological safety and follower prohibitive voice using matched Chinese leader–follower observations. Ito et al. (2026) reported, in a conference abstract, a vignette experiment linking leader work-error sharing, perceived psychological safety and follower willingness to share errors. Raychev (2026) integrated studies of silence and voice in AI-driven workplace learning, and Lin et al. (2026) found differing associations between algorithmic management functions and employee voice. Field evidence outside AI is stronger than often assumed. In 95 teams, Zhang, Zhao and Yin (2023a) found that members were more willing to share errors with leaders who openly discussed their own. McGuire, De Cremer and Narayanan (2026) moved closer to the AI setting, linking leader-provided opportunities to voice AI-related concerns, received with humility, to employees' trust in AI. None of these should be redescribed as a definitive test of whether a manager openly acknowledging AI-output errors predicts real employee warnings about those outputs. The distinction is the focal analytic problem, not a claim that the broader AI-and-voice literature is empty.

This review asks three questions. First, what does the available evidence imply about the transition from exposure to an AI-related problem to risk-directed employee voice? Second, how might managerial error disclosure and psychological safety influence this transition beyond generic leader openness? Third, how could monitoring, reporting channels and error opportunities condition or obscure the outcome? The contribution is a critical synthesis and an explicit measurement and research agenda for organisational behaviour and human-resource management. No original survey, experiment or effect-size meta-analysis is reported.

2. Methodology

2.1 Review design and epistemic scope

The study uses a critical integrative review design. It sets empirical studies, established organisational theories, adjacent systematic reviews and a few technical or governance sources side by side. The aim is to explain a problem that does not yet have a mature, shared measurement tradition. This is not presented as an exhaustive systematic review. The purpose is to evaluate conceptual fit, contradictions, transferability and missing links, rather than estimate pooled effects or claim that every eligible database record was captured. The unit of synthesis is a claim supported by a particular study design and construct, not a publication vote count. Studies are not treated as interchangeable merely because they contain the words "AI", "voice" or "safety".

The empirical target is an employee's upward or formally routed warning that a GenAI-assisted work output may cause inaccuracy, bias, privacy loss, unfair treatment or unsafe advice. Its object is an identifiable risk in the content or use of an output. Generic preferences about AI training, enthusiasm for technology and ordinary process-improvement suggestions are recognised as adjacent, not automatically equivalent outcomes. "Managerial AI error disclosure" refers to a supervisor visibly acknowledging that an AI-generated recommendation was incorrect or unsafe, describing verification or correction and declining to blame the person who noticed it. These operational descriptions are proposed by the review; they are not existing validated AI-specific scales.

2.2 Evidence search and selection

A bounded search was completed on 8 October 2026 using targeted scholarly web retrieval, publisher and DOI records, and citation-led discovery starting from the supplied research plan. Four query families guided the search: ("generative AI" OR "ChatGPT") AND ("employee voice" OR "organisational listening"); ("AI awareness" OR "algorithmic management" OR "workplace surveillance") AND ("voice" OR "speaking up"); ("psychological safety" OR "leader error sharing" OR "managerial openness") AND ("prohibitive voice" OR "error reporting"); and ("hallucination" OR "AI overreliance") AND ("human oversight" OR "correction"). Recent publications from 2022–2026 were prioritised for technological context, while earlier foundational studies were retained when necessary to define constructs and competing pathways. A supplementary benchmark search on 22 November 2026 targeted journals of six major publishers to identify recent studies that the first search had missed, and added nine publications (Appendix A5). Relevant English-language peer-reviewed research and substantive reviews were favoured; one explicitly identified conference abstract and one risk-governance report were retained with their evidence status made clear.

Candidate studies were screened for an identifiable bibliographic record and a substantive contribution to at least one of five themes: GenAI-output risk, voice form and opportunity, leader signals and safety, algorithmic controls, or methodological identification. DOI, bibliographic identity and stated study purpose were checked against journal, conference, scholarly-institution or official framework pages where accessible. For some paid publications only an abstract, highlights or bibliographic record was obtainable. Their results are described only at that level. Full-text material was consulted selectively for accessible studies; no claim is made that all references were independently read in full. An

item was excluded from evidentiary comparison when its topic was AI adoption but its construct could not inform reporting, or when source identity could not be corroborated. The search was purposive, not a registered database search; consequently there is no fabricated PRISMA flow, initial hit count, formal risk-of-bias score or exhaustive denominator.

2.3 Synthesis and evidence-strength rules

The synthesis followed three passes. First, studies were grouped by what they actually measured: general or promotive voice, prohibitive/safety voice, error communication, AI use and reliance, psychological safety, or algorithmic monitoring. Second, their designs were distinguished. Experiments can support intervention contrasts in their setting. Temporally separated surveys reveal time-ordered associations but generally not causal mediation. Cross-sectional and interview evidence can indicate plausible mechanisms but cannot identify directional effects, and theoretical reviews consolidate constructs without producing new field estimates. Third, apparent links were mapped into a process model, with a separate label for evidence-backed regularities and review-derived propositions.

When assessing the proposed AI-risk outcome, we gave most weight to direct operational correspondence, ahead of recency or journal reputation. For example, a study of GenAI adoption suggestions is direct evidence about AI-related suggestions but only indirect evidence about error reporting. A safety-voice meta-analysis gives broader background on reporting risks but does not estimate AI-specific voice. Thematic comparisons are summarised in Table 1. Because outcomes, designs and populations differ substantially, numerical synthesis or a single “overall effect” would be unjustified. Appendix A records the search logic and key boundary decisions so later researchers can expand the coverage into a registered systematic review.

3. Results: Thematic Evidence Synthesis

3.1 AI productivity and the need for corrective information

The literature establishes two facts that must be held together. GenAI assistance can improve the speed and assessed quality of particular tasks (Noy & Zhang, 2023; Brynjolfsson et al., 2025). Generated outputs can also contain hallucinated or risky content (Ji et al., 2023; Weidinger et al., 2022). The first fact creates organisational incentives for adoption; the second makes error detection and escalation relevant to responsible use. Neither evidence stream, however, establishes how frequently employees challenge a faulty output in routine organisational settings. The distribution of risks depends on task stakes, data access, model behaviour and verification practices. It would be a category error to infer workplace reporting rates from a technical hallucination benchmark.

Research on reliance further complicates the idea that providing an AI explanation ensures adequate human correction. Glikson and Woolley (2020) organise prior evidence on trust in AI, but trust is not necessarily calibrated to performance in a particular instance. In a human–AI decision experiment, Buçinca et al. (2021) found that cognitive-forcing interventions could reduce overreliance, while imposing usability costs; these findings concern individual interaction with decision support, not manager-directed voice. The conceptual implication is nevertheless consequential: a worker must first recognise a reason to doubt an output before it becomes meaningful to assess willingness to report. Silence cannot be diagnosed from the absence of reports when error detection or meaningful exposure is unobserved.

Human–AI complementarity has long been framed in terms of distinct strengths in organisational decision-making (Jarrahi, 2018; Raisch & Krakowski, 2021). A corrective-voice account specifies one social condition under which that complementarity might operate: employees must be able to transfer local, task-specific doubts to someone capable of modifying or halting use. Accordingly, a mature governance model should examine the complete pathway from the generated recommendation through human scrutiny to organisational handling, rather than viewing “a human in the loop” as an automatic safeguard. The National Institute of Standards and Technology framework supplies a risk-management rationale for governing, mapping, measuring and managing AI risks, but it does not experimentally validate an employee-voice intervention (Tabassi, 2023).

3.2 Risk voice is not adoption voice, and non-reporting is not always silence

Employee voice research distinguishes constructive proposals from the communication of potential harm. Liang et al. (2012) operationalised promotive and prohibitive voice as related but distinct behaviours; Van Dyne et al. (2003) emphasised differentiated forms of employee voice and silence. Morrison (2014, 2023) reviewed how the decision to speak up depends on expected consequences and utility, among other factors. These distinctions matter when machine-generated output is the object of concern. A worker recommending wider GenAI use may be promoting adoption, while the same worker's request to withdraw a biased AI-generated explanation is warning against harm. Both are constructive, but they should not automatically be pooled into a single measure.

Dong et al. (2024) provide particularly strong reasons to distinguish them. Their US survey evaluated organisational listening during GenAI training, perceived autonomy and competence, adoption attitudes and voicing of implementation suggestions. The reported positive pattern supports an adoption-voice pathway. It does not reveal whether these employees raised concerns about incorrect customer-facing responses, disclosed privacy problems or challenged

managers' preferred outputs. Lin and He (2024) likewise document connections between AI awareness and broader voice in hospitality, with mediating and moderating variables relevant to employees' appraisals of change. Zhou and Lyu (2025) similarly connect leaders' AI awareness to employee voice through hindrance and challenge stressors. These studies provide adjacent evidence about how AI context enters voice behaviour; their measures should not be silently relabelled as "AI-error whistleblowing".

A second distinction concerns exposure. For conventional voice measures, a low number of suggestions often stands in for lower speaking-up activity. For AI-risk voice, the opportunity to encounter a questionable output is a logical precondition. A worker handling many AI-generated client messages has a different reporting opportunity from one who uses GenAI only for harmless internal brainstorming. Recorded zero reports could indicate no defect, a defect not noticed, a noticed risk communicated through another channel, or an intentionally withheld warning. These states cannot be separated from a single agreement scale. Accordingly, reporting propensity should be analysed conditional on opportunity or accompanied by a parallel incident-exposure measure. That measurement recommendation is derived from the constructs; no included study yet establishes its superiority for this specific outcome.

Evidence on safety voice offers a productive but bounded comparator. The meta-analysis by Tedone et al. (2026) synthesised 49 independent studies and associated safety-supportive leadership and work climates with safety voice. Such findings make hazard reporting a plausible neighbouring framework, particularly where an erroneous AI output could affect customers or colleagues. They cannot supply an AI-specific effect estimate, and the stakes of physical workplace safety do not always coincide with risks of generated misinformation. Raychev's (2026) integrative review identifies voice and silence patterns during AI-driven workplace learning, including learning-oriented and defensive responses. Its focus on learning and reskilling complements, rather than replaces, the output-specific reporting problem addressed here. A recent concept review of 86 safety-voice studies makes the same point from the safety side. Research rarely specifies the hazards that prompt people to speak up, and it concentrates on verbal, reactive speaking up (Paul, Pilbeam, & Smallwood, 2025). An AI-risk voice construct that names the risky output and the opportunity to report it would avoid inheriting that weakness.

3.3 Managerial openness, acknowledgement of errors and psychological safety

Leader responses form a credible candidate mechanism because risk-oriented voice challenges an existing practice and can impose interpersonal costs. Detert and Burris (2007) linked leader openness to voice in a service setting, while Burris (2012) documented potential managerial penalties for challenging forms of voice. These studies establish that the reception of an employee message can matter as much as an abstract invitation to participate. Organisational error-management research similarly treats detection, communication and correction of errors as cultural practices (van Dyck et al., 2005). Yet none of these findings alone specifies the effect of acknowledging a particular GenAI output error.

Leader error disclosure should be defined behaviourally rather than as a flattering self-description. In an AI workflow, a manager might say that a generated answer was unsupported and show how the team identified the problem. The manager would then withdraw or revise the recommendation and confirm that the employee who flagged it will not be penalised. This differs from claiming that the manager is approachable or believes in innovation. Work on expressed humility suggests that acknowledging limitations can be related to learning-related processes, although those studies are not AI-specific (Owens et al., 2013). The closest directly relevant field evidence comes from Zhang, Zhao and Yin (2023a), who studied 353 employees in 95 teams from a social-learning perspective. Leaders' open discussion of their own errors was positively related to members' sharing of errors with the leader. A 2026 Academy of Management Proceedings abstract by Ito et al. reports a randomised leader-error-sharing vignette with 175 workers and a mediated willingness-to-share-errors outcome. Together these studies make the disclosure–sharing link plausible, but both concern employees' own mistakes rather than warnings about a machine-generated output, and the Ito et al. evidence is abstract-level and hypothetical. Error disclosure also carries costs for the person who speaks. Using a self-disclosure framework, the same authors show that error sharing can shape leaders' judgements of the employee's ability and integrity, depending on how visible and severe the error is (Zhang, Zhao, & Yin, 2023b).

Psychological safety offers an interpretable pathway connecting the signal to employee action. Edmondson (1999) established the concept in work-team learning, while Frazier et al. (2017) and Newman et al. (2017) synthesised evidence concerning its antecedents and correlates. A manager who names an AI mistake without blame may reduce perceived interpersonal risk associated with challenging the next output. In matched survey data from Chinese organisations, Tian et al. (2025) associated leader prohibitive voice with follower prohibitive voice through psychological safety. Humble leadership has likewise been linked to teams' ability to overcome silence (Zettna, Nguyen, Restubog, Schilpzand, & Johnson, 2025). Tian et al.'s result closely supports the plausibility of the pathway, but its exposure is leader prohibitive voice rather than manager acknowledgement of an AI-generated error. Treating it as direct proof of an AI-specific mechanism would overstate construct transportability.

Safety is also insufficient by itself. An employee may believe reporting will be welcomed yet lack the knowledge to identify an error, the authority to pause deployment or confidence that the warning will reach a decision-maker. Conversely, a technically competent employee may recognise a defect but choose not to report because managers appear invested in the system. These conditions separate the ability to detect from the choice to communicate and the

effectiveness of the receiving channel. The literature reviewed here supports the importance of openness and safety for voice in general; the specific incremental pathway from AI-error disclosure to risk-voice behaviour remains a testable integrative proposition.

3.4 Algorithmic management, monitoring and the credibility of invitations to speak

Monitoring is not a single exposure. Digital systems may schedule work, allocate tasks, provide developmental feedback, audit compliance or record individuals for sanctions. Kellogg et al. (2020) conceptualised algorithmic control as a contested field of organisational power; Keegan and Meijerink (2025) reviewed the broader employment implications of algorithmic management. A more recent review by Chen et al. (2026) identifies different configurations of workplace algorithmic management rather than assuming one homogeneous technology effect. This literature suggests that a visible AI log is not inherently punitive: an auditable correction trail can support accountability even as personnel surveillance can increase perceived risk.

The empirical direction of monitoring-related evidence varies with function and context. Kensbock and Stöckmann (2021), studying 100 employees in industries experiencing digital transformation, reported that perceived technological surveillance weakened learning and voice-related processes. Kayas (2023) documented fragmented and tension-filled workplace-surveillance literatures in a systematic review. More directly, Lin et al. (2026) analysed 351 matched employee–supervisor cases in a Chinese platform enterprise. Algorithmic feedback was associated with greater voice through felt responsibility for constructive change, whereas directing, scheduling and monitoring functions showed negative pathways. These findings do not isolate GenAI or interpersonal sanction for challenging AI errors, but they rule out an indiscriminate proposition that any algorithmic monitoring always suppresses all voice.

The theoretically important moderator is therefore punitive interpretation, not the mere presence of monitoring. A manager can acknowledge a tool's mistake and still undermine that signal if records of objections affect performance appraisal or workload allocations. Alternatively, a transparent, non-punitive incident register may make concerns easier to trace and resolve. What matters is whether the employee expects an AI-risk warning to carry personal costs, how the organisation separates quality assurance from discipline, and whether managers actually respond consistently with their declarations. These are hypotheses about the credibility of managerial communication, informed by adjacent research rather than established causal effects in GenAI teams.

3.5 Evidence contrasts and limits of transfer

Three evidence contrasts are especially informative. First, GenAI adoption studies describe employee suggestions and positive attitudes (Dong et al., 2024; Lin & He, 2024), whereas prohibitive-voice theory concerns communicating harmful practices (Liang et al., 2012). Their outcomes should be disaggregated. Second, leader and safety research establishes general relationships (Detert & Burris, 2007; Tian et al., 2025). The closest leader-error-sharing evidence, a team-level field study and a vignette experiment, concerns employees' own errors (Ito et al., 2026; Zhang et al., 2023a). Whether the same signal transfers to warnings about AI-generated outputs under real workplace incentives remains uncertain. Third, algorithmic management evidence includes both inhibitory monitoring and enabling feedback (Lin et al., 2026); consequently a one-dimensional "AI control" index may suppress function-specific differences.

Existing reviews already cover substantial adjacent terrain. Morrison (2023) surveys employee voice and silence generally; Guarin et al. (2025) review longitudinal voice research; Raychev (2026) centres AI-driven workplace learning; Liu et al. (2026) examine psychological mechanisms of conversational AI; and Chen et al. (2026) classify algorithmic-management configurations. The present contribution is narrower. It aligns four elements: the object of voice (a risky GenAI output), the trigger (an encountered incident), the managerial signal (error acknowledgement) and the response (an observable warning routed for action). This synthesis should not be confused with a new pooled empirical effect, nor with a demonstration that earlier scholars have never considered error correction.

Table 1. Evidence map of selected direct and adjacent studies

Source and class	Design or corpus	Finding relevant to synthesis	Reason for limited transfer
Dong et al. (2024) [D]	US employees; survey (n = 500)	Listening during GenAI training related to adoption suggestions via autonomy, competence and attitudes	Outcome is adoption-oriented voice; not a count of error warnings
Lin & He (2024) [D]	Hospitality; cross-level survey	AI awareness associated with general employee voice through contextual pathways	Voice about errors in specific outputs was not isolated
McGuire et al. (2026) [D]	Empirical study of employees (abstract-level use)	Leader-provided voice opportunities linked to trust in AI through reduced negative affect; humility strengthened the indirect path	Outcome is trust in AI, not reporting of AI-output errors
Tian et al. (2025) [A]	China; 302 matched dyads	Leader prohibitive voice and follower voice associated through psychological safety	Leader behaviour is not GenAI-error disclosure
Ito et al. (2026) [A]	Vignette experiment, n = 175; conference abstract	Leader error sharing linked with follower error sharing via perceived safety	Abstract only; self-reported hypothetical willingness
Zhang et al. (2023a) [A]	95 teams, 353 employees; field survey	Leaders' open discussion of own errors positively related to members' error sharing with leaders	Employees' own errors; no AI-output or risk-voice outcome
Lin et al. (2026) [D/A]	China; 351 employee–supervisor cases	Feedback related positively to voice; directing, scheduling and monitoring negatively	Algorithmic platform management, not GenAI-output reporting
Kensbock & Stöckmann (2021) [A]	Survey, n = 100	Technology surveillance attenuated learning/voice-related process	Digital transformation, not GenAI-specific
Raychev (2026) [R]	Integrative review; 61 retained studies	Frames silence and voice in AI-driven workplace learning	Learning rather than AI-output incident reporting
Tedone et al. (2026) [R]	Meta-analysis; 49 independent studies	Safety-supportive contexts associated with safety voice	Hazard voice broader than GenAI-generated-content risks
Chen et al. (2026) [R]	Systematic review; 167 studies	Differentiates algorithmic-management configurations	Does not identify the proposed disclosure-to-risk-voice causal path

4. An Integrative Model of Corrective AI-Risk Voice

4.1 A five-stage reporting process

The review yields a process account with five analytically separable stages. Stage 1, risk exposure, concerns the presence and consequence of a questionable GenAI output in an employee's work. Stage 2, detection, requires the employee to notice an inconsistency or hazard; competence, workflow design and AI reliance may influence this step. Stage 3, appraisal, involves weighing uncertainty, seriousness, expected voice efficacy and possible interpersonal costs. Stage 4, reporting, is a message directed to a supervisor or authorised risk channel, including any decision to escalate rather than remain silent. Stage 5, organisational response, covers verification, correction, feedback and any sanctions or learning. Each stage can fail even when preceding stages function properly. Figure 1 sets out the sequence.

This separation corrects a recurrent construct problem. Organisations cannot infer a low willingness to report from a low raw number of warnings when few risk opportunities occurred, detection was weak or warnings went to a different channel. Equally, high observed reporting may indicate a transparent culture, a high defect burden or both. For this reason the proposed main quantity is an opportunity-conditioned probability or rate of a substantiated warning, reported alongside error incidence, detection and channel usage. "Substantiated" here means reviewed against an agreed risk category, not automatically verified as a genuine model defect. In most settings, multiple reviewers or incident audits will be needed to distinguish reasonable concern from ultimately confirmed error.

Managerial error acknowledgement is located primarily between appraisal and reporting: it may signal that challenge is legitimate and will not be punished. It can also shape detection when managers actively demonstrate verification. Psychological safety is an individual perception or team property only when aggregation is justified, not an automatically interchangeable variable across analytic levels (Edmondson, 1999; LeBreton & Senter, 2008). Monitoring operates across the cycle: it can supply incident records, influence the credibility of a non-retaliation message and alter the costs attached to escalation. Organisational response closes the loop and may affect whether subsequent warnings are raised, but that dynamic feedback requires genuinely repeated observations to test (Guarin et al., 2025).

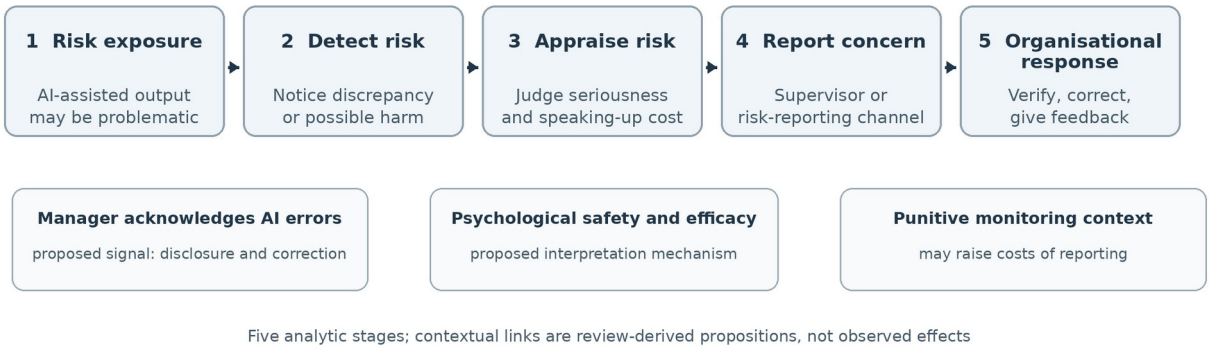

Figure 1. Review-derived stages of corrective voice about GenAI output risks

4.2 Review-derived propositions

P1 (specificity). Manager acknowledgement of a documented GenAI-output mistake will be more closely associated with AI-risk prohibitive voice than with generic pro-adoption suggestions, once broad leader openness and baseline voice are considered. A failure to distinguish the outcomes would imply that the signal principally reflects general supportive leadership rather than an AI-correction mechanism.

P2 (interpretation). Perceived psychological safety will partially account for the time-ordered relationship between managerial acknowledgement and subsequent reports of AI-related risks, but evidence from non-randomised surveys would warrant an associational, not causal, mediation interpretation. A competing explanatory route through competence or felt responsibility for constructive change should be modelled explicitly (Lin et al., 2026).

P3 (sanction credibility). Expected punishment or adverse appraisal following AI-related challenge will attenuate any positive association between managerial acknowledgement and risk voice. The predicted condition is punitive use of monitoring data; routine and transparent quality-assurance auditing is not presumed to have the same direction.

P4 (opportunity dependence). The disclosure–voice association will be better estimated after classifying opportunities to detect and report material AI risks than from unadjusted frequency of warnings alone. Without observation of opportunities, a low warning rate mixes absence of faults, missed detection and withholding of concerns.

P5 (response feedback). A manager's verified, timely and non-punitive handling of prior AI warnings may be a stronger predictor of future risk voice than verbal disclosure without corrective follow-through. This proposition extends the process into repeated organisational learning and is not a result observed in the reviewed GenAI-specific studies.

The propositions serve as discriminating tests, not as claims that the review itself has confirmed them. They preserve rival explanations and set boundaries around what would count as novel empirical evidence.

4.3 The nearest alternatives

One rival explanation is generic leader quality: managers who disclose AI errors may already have better relationships, stronger trust norms or lower employee turnover. A second is task exposure: teams using GenAI in more demanding settings will have more questionable outputs and more occasions to report. A third is AI competence: knowledgeable employees may identify and articulate risks independently of the managerial signal. A fourth is reporting-channel selection: employees who distrust a line manager may communicate to compliance officers, peers or informal forums, disappearing from supervisor-rated measures. A further rival is disclosure cost. Error sharing can lower a leader’s assessment of the sharer’s ability or integrity (Zhang et al., 2023b), so employees may weigh reputational risk independently of any safety signal. A fifth is managerial favouritism: supervisors may remember warnings from liked employees and disregard others. The integrative model must therefore distinguish who encounters a risk, who detects it, where it is reported and who evaluates the report.

5. Research Agenda and Methodological Priorities

5.1 Measure the outcome as behaviour under opportunity

The immediate priority is an AI-specific incident taxonomy distinguishing hallucination or unsupported factual assertion, discriminatory or inappropriate recommendation, privacy or data-governance breach, and safety-sensitive advice. Employees should be asked about exposure to these categories without soliciting actual client data or proprietary prompts. A parallel item should ask whether the concern was reported and to whom, with a distinct "no opportunity" state. Supervisor ratings can assess observable warnings but should not be considered a complete event registry. Where organisational permissions allow, de-identified ticket or quality-assurance data could triangulate frequency, routing and follow-up without converting sensitive work records into unrestricted research material.

Construct-validation work must establish separability. Manager AI-error acknowledgement should be measured by concrete prior actions, not a generic "my manager supports me" statement. Psychological safety should be distinguishable from leader openness and job security. Punitive monitoring should be distinguished from legitimate auditing and performance feedback, and AI-risk voice from promotive voice and informal dissatisfaction. Cognitive interviews, bilingual translation checks where relevant and confirmatory measurement models could examine these distinctions. Baseline safety, voice, competence and GenAI use should be recorded before subsequent outcome measurements.

5.2 Identify a credible comparison and temporal sequence

A three-wave employee–supervisor panel is a feasible starting point in Malaysian service organisations, reflecting the original research plan, but it cannot identify a causal mechanism on its own. Wave 1 could measure prior leader behaviour, general openness, exposure and baseline voice; Wave 2 psychological safety and opportunity; and Wave 3 externally rated AI-risk voice, separated from generic suggestions. Multilevel estimates should account for team nesting, unequal team sizes and individual-versus-team variance; justified mediation models should avoid combining within-team and between-team pathways indiscriminately (Preacher et al., 2010). The mere separation of survey waves does not remove confounding or reverse causation.

A stronger complementary design would randomise a managerial response script or ethically approved training intervention at the team level. One condition would model transparent acknowledgement and non-punitive correction of an actual or simulated AI failure; comparison conditions might offer general leader encouragement or neutral verification instructions. Subsequent outcome opportunities should be standardised so that groups encounter comparable AI-risk scenarios. Randomised vignette studies can test immediate perceptions, as Ito et al. (2026) exemplify in general leader-error sharing, but studies of real workplace behaviour are still needed for external validity. Research on AI overreliance suggests that detection-support tools could be manipulated separately from managerial messaging (Buçinca et al., 2021).

Power analysis should reflect the number of teams, intraclass correlation, event rarity and the effective number of exposed workers rather than a simple nominal employee total. If natural incident rates are too low, planned observation can be extended or a labelled behavioural simulation can be added, but simulation outcomes must not be passed off as actual organisational warnings. Preregistered exclusions, alternative outcome definitions, attrition handling and sensitivity models are necessary. Multiple measurement sources reduce certain shared-source biases but cannot eliminate selection, memory effects or supervisor preferences (Podsakoff et al., 2003).

5.3 Test contexts and boundary mechanisms

Future studies should compare authorised and informal GenAI use, task-critical and low-stakes outputs, and line-manager versus independent escalation channels. The relevant mechanism may differ between a customer-service draft and a compliance-sensitive recommendation; privacy constraints could also limit what a worker can safely disclose about the underlying prompt. The technology's reliability and the employee's opportunity to inspect source material may change detection more than willingness to voice. Culture, employment security and organisational history may alter perceived interpersonal costs, so evidence from Chinese platform organisations or US GenAI-training samples should not be assumed to transfer unmodified to Malaysian service teams.

Channel design deserves experimental attention in its own right. In two experiments, technical anonymity increased employees' willingness to speak up through psychological safety (Liu, Yuan, & Luo, 2025); an anonymous AI-incident channel could therefore be compared with a named line-manager route. Monitoring research should separate supportive feedback, quality-control audit, performance ranking and perceived punitive surveillance. The differential findings of Lin et al. (2026) make a functional decomposition preferable to a single omnibus technology-control score. The institutional meaning of the correction channel should also be recorded. A manager's promise of openness can be undermined if workers see negative appraisal consequences, whereas credible protected escalation may preserve reporting even when the direct manager is reluctant. Repeated observations after real management responses would help test whether the organisation learns from the warning and whether future voice changes. Table 2 lists the discriminating tests that follow from the propositions.

Table 2. Discriminating empirical tests derived from the review

Proposition	Priority comparison or measurement	Possible result challenging the proposition
P1: signal specificity	Manager AI-error disclosure vs general leader openness; AI-risk voice vs adoption suggestions	No incremental association with AI-risk voice after controlling openness
P2: safety pathway	Wave 1 behaviour, baseline safety, Wave 2 safety, Wave 3 routed risk warnings	Safety association disappears with baseline conditions or channel measurement
P3: punitive boundary	Separate sanction risk, compliance audit, scheduling and developmental feedback	No differential association across punitive monitoring conditions
P4: reporting opportunity	Risk events, detected events, channel and independently assessed warning	Simple reporting-rate differences explained by exposure and detection
P5: response feedback	Recorded correction quality and later validated risk warnings	Disclosure without corrective action predicts equally well as verified follow-through

6. Discussion

6.1 Contribution to organisational behaviour and GenAI governance

The primary contribution of this review is not another general claim that psychological safety predicts employee voice. That relationship is extensively theorised and supported in adjacent settings (Edmondson, 1999; Frazier et al., 2017; Morrison, 2023). The new integrative focus is the narrower, more actionable transition from a risky AI-generated output to a routed corrective warning. This transition demands the joint presence of an error opportunity, human detection, a credible communication pathway and appropriate managerial response. By placing reporting opportunity and channel visibility inside the theory, the model addresses a weakness of raw voice counts. The same low count can reflect either reliable outputs or a failure to surface risks.

The review also shows why managers and algorithmic systems must be considered together. Organisational leaders communicate how to respond to errors, while algorithmic monitoring functions can reinforce or contradict those messages. Generalised claims that technology increases or decreases voice overlook the contradictory signals supplied by different features of algorithmic management (Kensbock & Stöckmann, 2021; Lin et al., 2026). Research on conversational AI and worker experience expands the psychological context (Liu et al., 2026), but it should be connected to a behaviourally precise criterion when the practical concern is error containment. In this account, GenAI governance has an epistemic problem (whether an error is noticed) and a social problem (whether the notice becomes an actionable report).

6.2 Conditional managerial implications

Organisations can use the process model as a diagnostic checklist without treating its untested propositions as proven interventions. First, they can define what counts as a reportable AI-related problem and provide a safe, privacy-preserving routing channel. Second, they can distinguish a verified correction from an employee's reasonable warning, allowing reporting even before the concern is fully established. Third, leaders can make corrections visible and explain the decision process without disclosing confidential prompts. Fourth, governance and HR teams can review whether performance monitoring creates adverse incentives for reporting. These suggestions are consistent with general error-management and risk-governance principles (van Dyck et al., 2005; Tabassi, 2023), but direct evidence of their effects on AI-risk voice remains to be generated.

Leaders' own learning matters as well. Asking leaders to recall what they learned from past mistakes increased their expressed humility (Hu, Zhang, Lount, & Tepper, 2024). This suggests a low-cost way to prepare managers to acknowledge AI errors credibly. Because speaking up has consequences for voicers, co-workers and managers alike (Nieberle & Fladerer, 2025), the organisation's visible response to the first warnings is likely to shape whether later warnings occur. Training alone is unlikely to answer every part of the problem. Better AI literacy could increase detection, while a punitive climate could still discourage reporting. Conversely, a friendly environment could encourage unverified complaints unless staff have a transparent triage process. A balanced design therefore combines detection capability, non-retaliation, feedback, traceability and independent review in proportion to the stakes of a particular use case. The effectiveness of this package should be tested rather than asserted from adjacent voice studies.

6.3 Limitations of the review

This is a critical integrative review with a bounded and selective evidence search; it is not a registered systematic review or meta-analysis. Search interfaces did not yield a complete, exportable database universe, and no claim of exhaustive coverage is made. Several directly relevant papers were accessible only through abstracts or publisher summaries, particularly the 2026 leader-error vignette conference item and some subscription research. These were explicitly limited in the inference they support. Publication timing also matters: articles newly available in 2026 may not yet have independent replication. No quantitative risk-of-bias grading or pooled effect is claimed.

There are also conceptual limitations. AI-risk voice is defined for the purpose of this synthesis and lacks an established standardised scale; workplace risk classes can overlap, and not all AI errors warrant supervisor escalation. The argument transfers selectively from safety voice, general organisational voice, leader humility and algorithmic management; mechanisms established in these domains may not generalise to GenAI contexts. Proposed propositions could fail after adjusting for existing trust climates, competence, incident prevalence and reporting routes. Those are meaningful potential findings, not problems to conceal. The review accordingly offers a testable framework and carefully bounded synthesis rather than empirical proof of a new management intervention.

7. Conclusion

GenAI deployment makes employee correction of risky outputs an organisational problem rather than solely a technical quality-control task. The available literature supports attention to voice type, managerial response, psychological safety and the differing functions of algorithmic monitoring. It does not yet establish that a supervisor's admission of an AI-output error causes later employee risk reporting. The integrative framework developed here distinguishes exposure, detection, appraisal, routed voice and organisational response, and it specifies where unsupported assumptions are often introduced. Its most useful implication for future research is methodological: measure reportable risk opportunities and reporting channels alongside psychological and managerial constructs. That approach could determine when human oversight becomes an observed corrective practice rather than an aspiration attached to AI adoption.

Acknowledgement

Funding and acknowledgements require confirmation by the accountable author(s) before submission. No sponsor or institutional contribution is asserted in this manuscript.

Research Ethics, Data Availability and AI-Assistance Disclosure

This review did not recruit human participants or analyse private participant-level data. No new empirical dataset was generated. A source list, dated search strings and evidence-scope notes are reproduced in the manuscript and Appendix A; copyrighted full texts are not redistributed. An AI writing assistant supported the retrieval, drafting, synthesis and document formatting process. Before any submission, the human author(s) must independently verify bibliographic records and substantive claims, confirm author contributions and funding disclosures, and approve any journal-specific AI-use statement.

Conflict of Interest

Author conflict-of-interest declarations have not been provided and must be completed by the accountable author(s). No declaration of their interests is made on their behalf.

Appendix A: Search and Synthesis Protocol

A1. Search date: 8 October 2026. Sources consulted: targeted scholarly web searches, accessible publisher and DOI records, official institutional records, citation tracing and the supplied 2026 research plan. The search was structured but not database-exhaustive; it must not be labelled a PRISMA systematic review.

A2. Search families: (1) “generative AI”/“ChatGPT” + “employee voice”/“organisational listening”; (2) “AI awareness”/“algorithmic management”/“workplace surveillance” + “voice”/“speaking up”; (3) “psychological safety”/“leader error sharing”/“managerial openness” + “prohibitive voice”/“error reporting”; (4) “hallucination”/“AI overreliance” + “human oversight”/“correction”. Search terms are conceptual variants, not a literal common Boolean query across databases.

A3. Inclusion logic: verified scholarly identity plus direct or mechanism-relevant findings; seminal works retained for construct definitions; 2022–2026 sources prioritised for AI-related contexts. Exclusions: unverifiable citations, purely promotional material and papers not informing risk detection, voice, leader signals, monitoring or their research design. No automated bibliometric completeness, PRISMA count or independent full-text screening by multiple reviewers is claimed.

A4. Source interpretation categories: D = direct AI-and-voice evidence; A = adjacent field evidence (voice, safety, error communication, monitoring); T = technical or governance context; R = review or meta-analysis. Abstract-only

evidence is not treated as full-text verification. Table 1 displays selected comparators rather than an exhaustive study-by-study risk-of-bias matrix.

A5. Supplementary benchmark search (22 November 2026): targeted searches of journals published by Elsevier, Wiley, Springer Nature, Sage, Emerald and Taylor & Francis combined ("leader humility" OR "leader error sharing" OR "admitting mistakes") with ("employee voice" OR "error reporting"), and ("generative AI" OR "artificial intelligence") with ("speaking up" OR "voice"). Nine publications were added; all bibliographic records in the reference list were re-resolved against Crossref on the same date.

A6. Future update route: run documented searches on Scopus, Web of Science, Business Source Complete and OpenAlex where licensed; export complete retrieval sets; and record database- and date-specific query strings, duplicates and screening outcomes. Screening should be independent. Criteria should then be registered, and a formal systematic review or meta-analysis considered only if enough studies with comparable outcome measures exist.

References

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 188. https://doi.org/10.1145/3449287

Burris, E. R. (2012). The risks and rewards of speaking up: Managerial responses to employee voice. *Academy of Management Journal, 55*(4), 851–875. https://doi.org/10.5465/amj.2010.0562

Chen, W., Li, A., Pan, C., Yu, T., Ali, A., & Sun, Y. (2026). Algorithmic management in the workplace: A systematic review and topic modeling integration using BERTopic. *International Journal of Information Management, 86*, Article 102994. https://doi.org/10.1016/j.ijinfomgt.2025.102994

Detert, J. R., & Burris, E. R. (2007). Leadership behavior and employee voice: Is the door really open? Academy of Management Journal, 50(4), 869–884. https://doi.org/10.5465/amj.2007.26279183

Dong, E., Liu, H., Li, J.-Y., & Lee, Y. (2024). Motivating employee voicing behavior in optimizing workplace generative AI adoption: The role of organizational listening. *Public Relations Review, 50*(5), Article 102509. https://doi.org/10.1016/j.pubrev.2024.102509

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly, 44*(2), 350–383. https://doi.org/10.2307/2666999

Frazier, M. L., Fainshmidt, S., Klinger, R. L., Pezeshkan, A., & Vracheva, V. (2017). Psychological safety: A meta-analytic review and extension. *Personnel Psychology, 70*(1), 113–165. https://doi.org/10.1111/peps.12183

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Guarin, A. D., Townsend, K., Wilkinson, A., & Edwards, M. (2025). Time to voice? A review and agenda for longitudinal employee voice research. *Human Resource Management Review, 35*(1), Article 101059. https://doi.org/10.1016/j.hrmr.2024.101059

Hu, J., Zhang, S., Lount, R. B., & Tepper, B. J. (2024). When leaders heed the lessons of mistakes: Linking leaders' recall of learning from mistakes to expressed humility. *Personnel Psychology, 77*(2), 683–712. https://doi.org/10.1111/peps.12570

Ito, A., Chaiamarit, N., & Cuconato, G. (2026). When leaders admit errors: How psychological safety explains followers' willingness to share errors [Conference abstract]. *Academy of Management Proceedings, 2026*(1). https://doi.org/10.5465/AMPROC.2026.11061abstract

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human–AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Kayas, O. G. (2023). Workplace surveillance: A systematic review, integrative framework, and research agenda. *Journal of Business Research, 168*, Article 114212. https://doi.org/10.1016/j.jbusres.2023.114212

Keegan, A., & Meijerink, J. (2025). Algorithmic management in organizations? From edge case to center stage. *Annual Review of Organizational Psychology and Organizational Behavior, 12*, 395–422. https://doi.org/10.1146/annurev-orgpsych-110622-070928

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. https://doi.org/10.5465/annals.2018.0174

Kensbock, J. M., & Stöckmann, C. (2021). “Big brother is watching you”: Surveillance via technology undermines employees’ learning and voice behavior during digital transformation. *Journal of Business Economics, 91*, 565–594. https://doi.org/10.1007/s11573-020-01012-x

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Liang, J., Farh, C. I. C., & Farh, J.-L. (2012). Psychological antecedents of promotive and prohibitive voice: A two-wave examination. *Academy of Management Journal, 55*(1), 71–92. https://doi.org/10.5465/amj.2010.0176

Lin, Q., & He, L. (2024). Does artificial intelligence (AI) awareness affect employees in giving a voice to their organization? A cross-level model. *International Journal of Hospitality Management, 123*, Article 103947. https://doi.org/10.1016/j.ijhm.2024.103947

Lin, W., Zhang, M., Zhang, W., & Zhang, C. (2026). Will employees still speak up under algorithmic management? The differential effects of distinct algorithmic functions—Evidence from the Meituan platform in China. *Systems, 14*(5), Article 569. https://doi.org/10.3390/systems14050569

Liu, C., Yuan, Q., & Luo, J. (2025). Technical anonymity and employees’ willingness to speak up: Influences of voice solicitation, general timeliness, and psychological safety. *Management Communication Quarterly, 39*(2), 322–358. https://doi.org/10.1177/08933189241284598

Liu, L., Zhang, R., & Su, X. (2026). From tool to social actor: A systematic review of the psychological mechanisms through which conversational AI reshapes the employee experience. *Behavioral Sciences, 16*(9), Article 1704. https://doi.org/10.3390/bs16091704

McGuire, J., De Cremer, D., & Narayanan, D. (2026). How leaders build employee trust in artificial intelligence: Voice opportunities, humility, and trust transfer. *Journal of Business Research, 217*, 116470. https://doi.org/10.1016/j.jbusres.2026.116470

Morrison, E. W. (2014). Employee voice and silence. *Annual Review of Organizational Psychology and Organizational Behavior, 1*, 173–197. https://doi.org/10.1146/annurev-orgpsych-031413-091328

Morrison, E. W. (2023). Employee voice and silence: Taking stock a decade later. *Annual Review of Organizational Psychology and Organizational Behavior, 10*, 79–107. https://doi.org/10.1146/annurev-orgpsych-120920-054654

Newman, A., Donohue, R., & Eva, N. (2017). Psychological safety: A systematic review of the literature. *Human Resource Management Review, 27*(3), 521–535. https://doi.org/10.1016/j.hrmr.2017.01.001

Nieberle, K. W., & Fladerer, M. P. (2025). Voice echoes: A critical review and future research agenda on the consequences of employee voice for voicers, their co-workers, and managers. *Organizational Psychology Review, 15*(2), 209–255. https://doi.org/10.1177/20413866251317929

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Owens, B. P., Johnson, M. D., & Mitchell, T. R. (2013). Expressed humility in organizations: Implications for performance, teams, and leadership. *Organization Science, 24*(5), 1517–1538. https://doi.org/10.1287/orsc.1120.0795

Paul, J., Pilbeam, C., & Smallwood, A. (2025). Safety voice concept clean-up: Examining the voice that challenges us to be safer. *Safety Science, 191*, 106931. https://doi.org/10.1016/j.ssci.2025.106931

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*(3), 209–233. https://doi.org/10.1037/a0020141

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review, 46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Raychev, S. (2026). Employee silence and voice in AI-driven workplace learning: An integrative literature review. European Journal of Training and Development. Advance online publication. https://doi.org/10.1108/EJTD-05-2026-0142

Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Tedone, A. M., Mesmer-Magnus, J., Viswesvaran, C., & Newkirk, E. (2026). See something, say something: A meta-analytic review of safety voice. *International Journal of Management Reviews, 28*(1), Article e12401. https://doi.org/10.1111/ijmr.12401

Tian, X., Chae, H., Song, D., & Zhang, Y. (2025). Leader prohibitive voice shapes employee voice through psychological safety moderated by self-efficacy and generational differences. *Scientific Reports, 15*, Article 31469. https://doi.org/10.1038/s41598-025-17500-5

van Dyck, C., Frese, M., Baer, M., & Sonnentag, S. (2005). Organizational error management culture and its impact on performance: A two-study replication. *Journal of Applied Psychology, 90*(6), 1228–1240. https://doi.org/10.1037/0021-9010.90.6.1228

Van Dyne, L., Ang, S., & Botero, I. C. (2003). Conceptualizing employee silence and employee voice as multidimensional constructs. *Journal of Management Studies, 40*(6), 1359–1392. https://doi.org/10.1111/1467-6486.00384

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., . . . Gabriel, I. (2022). Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214–229). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533088

Zettna, N., Nguyen, H., Restubog, S. L. D., Schilpzand, P., & Johnson, A. (2025). How teams can overcome silence: The roles of humble leadership and team commitment. *Personnel Psychology, 78*(1), 67–102. https://doi.org/10.1111/peps.12660

Zhang, K., Zhao, B., & Yin, K. (2023a). When leaders acknowledge their own errors, will employees follow suit? A social learning perspective. *Journal of Business Ethics, 189*(2), 403–421. https://doi.org/10.1007/s10551-023-05329-9

Zhang, K., Zhao, B., & Yin, K. (2023b). The double-edged sword of error sharing in organizations: From a self-disclosure perspective. *Journal of Management Studies*. Advance online publication. https://doi.org/10.1111/joms.13003

Zhou, Y., & Lyu, B. (2025). How does leadership AI awareness shape employee voice behavior? Based on the framework of hindrance and challenge stressors. *Work, 82*(1), 289–305. https://doi.org/10.1177/10519815251341816