



AI-MEDIATED WELLBEING SURVEILLANCE AND THE BOUNDARY OF PSYCHOLOGICAL SAFETY

Dr. N Subbukrishna Sastry¹, Dr. Manjula Mallya M²

¹Professor, School of Management, CMR University, Bangalore, Karnataka, India.

ORCID ID: 0009-0009-0460-7057

²Associate Professor & Head, Dept. of Economics, Government First Grade College for Women Balmatta Mangalore, Karnataka, India. ORCID ID: 0009-0005-8812-6912

ABSTRACT

Employers increasingly deploy artificial-intelligence systems that read employee sentiment from written communication, infer exhaustion from calendar and workload digital exhaust, and flag individuals as at elevated risk of burnout — tools marketed under the language of care rather than control. This paper develops a conceptual model of how AI-mediated wellbeing monitoring reshapes psychological safety, the shared belief that a work environment is safe for interpersonal risk-taking on which voice, dissent and error reporting depend. Integrating psychological-safety theory, the algorithmic-management and workplace-surveillance literatures, and organisational voice research, the paper argues that wellbeing-monitoring intensity depresses psychological safety principally through a mediating perception of surveillance threat, and that this relationship is not fixed but is conditioned by the transparency of the monitoring system, the degree to which employees attribute developmental rather than punitive intent to it, and whether a human manager mediates between algorithmic output and managerial action. Five propositions specify these relationships, including a boundary condition for India's IT-BPM sector, where the Digital Personal Data Protection Act, 2023 supplies a general consent-and-purpose-limitation framework but no sector-specific rule for wellbeing analytics, leaving adoption norms to be set informally inside a high-attrition, always-on operating environment. The paper closes with implications for the design of psychologically safe monitoring architectures and an agenda for empirical validation.

KEYWORDS: Psychological Safety; AI-Mediated Monitoring; Sentiment Analytics; Burnout Prediction; Algorithmic Management; Employee Voice; Workplace Surveillance; India IT-BPM

1. INTRODUCTION

The instruments an organisation uses to know how its people are doing have changed faster than the theory available to interpret them. A decade ago, “knowing how employees are doing” meant an annual engagement survey and a manager's informal read of a team. Today it increasingly means a continuous, machine-generated signal: sentiment classifiers scoring the tone of email and chat traffic, workload-analytics dashboards that infer overload from calendar density and after-hours activity, and burnout-prediction models that combine these signals into an individual risk score an HR business partner or line manager can act on. These tools are typically introduced under an explicitly caring rationale — to catch exhaustion before it becomes attrition or a safety incident, to route support to the employees who most need it, to make wellbeing something the organisation manages proactively rather than discovers after the fact.

That caring rationale sits uneasily beside what the tools structurally are: a continuous, often opaque, individually resolvable stream of behavioural inference running beneath the employee's own awareness of what is being inferred and to whom it is reported. The wellbeing-monitoring literature and the workplace-surveillance literature have until recently developed largely in parallel — one asking whether these tools help employees, the other asking what continuous monitoring does to the monitored, largely in the context of productivity-tracking electronic performance monitoring rather than wellbeing-framed monitoring specifically. This paper treats that separation as itself the theoretical problem: the same infrastructure that promises earlier, better-targeted support for employee wellbeing is, from the employee's side of the relationship, also a surveillance apparatus, and the paper's central claim is that its effect on psychological safety — employees' willingness to voice dissent, flag problems, or admit difficulty — depends on which side of that description comes to dominate employees' lived experience of being monitored.

Psychological safety is the connective construct because it is precisely the belief structure that continuous, individually attributable monitoring is positioned to erode. Edmondson's (1999) foundational formulation defines psychological safety as a shared belief that a team is safe for interpersonal risk-taking — that raising a concern, admitting a mistake, or asking an awkward question will not be held against the person who raises it. Voice and error reporting, the behaviours through which organisations learn and correct course, are downstream of that belief (Detert & Edmondson, 2011; Morrison & Milliken, 2000). If AI-mediated wellbeing monitoring signals to employees that their internal state — their stress, their exhaustion, the tone of what they write — is itself continuously legible to the organisation, the paper argues this can plausibly narrow rather than widen the space employees judge safe for candour, even where the stated purpose of the monitoring is to help them.

The paper proceeds in five parts. Section 2 reviews psychological-safety theory, algorithmic management and workplace-surveillance research, and organisational-voice theory, weaving in the author's own published and in-preparation work on AI readiness, algorithmic transparency, and delegation as points of theoretical continuity. Section 3 develops a conceptual framework linking monitoring intensity to psychological safety through perceived surveillance threat. Section 4 advances five propositions,



three specifying moderators of the core relationship and one specifying a contextual boundary condition for India's IT-BPM sector. Section 5 assembles illustrative evidence consistent with the framework, and Section 6 examines the India application in depth before the paper closes with theoretical and managerial implications, limitations, and directions for empirical work.

2. LITERATURE REVIEW

This review draws on four literatures that have rarely been read together: the psychological-safety and organisational-voice tradition, which specifies what candour inside organisations depends on; the algorithmic-management and electronic-performance-monitoring tradition, which specifies what continuous machine oversight does to the people under it; the workplace-privacy and surveillance tradition, which supplies the boundary conditions under which monitoring is experienced as legitimate rather than intrusive; and a smaller, more recent literature on AI-mediated wellbeing tools specifically, where the caring framing is explicit and the surveillance framing is usually left implicit. The final subsection turns to India's IT-BPM sector, where wellbeing-analytics adoption is proceeding inside a distinctive regulatory and labour-market setting.

2.1 Psychological Safety as a Team- and Organisation-Level Construct

Edmondson (1999) defines team psychological safety as a shared belief, held in common by members of a team, that the team is safe for interpersonal risk-taking; it is a property of the collective climate rather than a personality trait of any one member, which is why it can be moved by organisational practices — including monitoring practices — rather than only by individual disposition. Edmondson and Lei (2014), reviewing the construct's first two decades of empirical life, trace its consistent downstream association with learning behaviour, error reporting, and voice, and note that the mechanisms through which organisational practices raise or lower psychological safety remain less well specified than its consequences, a gap this paper's monitoring-focused model is positioned to help close.

Two literatures specify what psychological safety governs. Morrison and Milliken (2000) describe organisational silence as a self-reinforcing collective phenomenon in which employees withhold concerns not because they lack them but because they judge the environment unsafe for raising them, producing a pluralistic ignorance in which management systematically under-hears the problems present in its own organisation. Detert and Edmondson (2011) extend this by showing that employees carry implicit theories — tacit, widely shared beliefs about what is and is not safe to say to whom — that operate independently of formal open-door policy and are calibrated from ambient organisational signals rather than explicit rules. AI-mediated wellbeing monitoring is, on this reading, exactly the kind of ambient signal from which implicit voice theories are built: employees infer what the organisation is safe to hear not from its stated policy on the monitoring tool but from what they believe the tool is actually capable of surfacing about them.

The theoretical stakes of extending psychological-safety theory to AI-mediated monitoring are therefore not narrow. If wellbeing analytics function as one of the ambient signals from which employees construct their implicit voice theories, then the design choices embedded in a monitoring system — what it observes, who sees the output, and how visibly its stated purpose is enforced — become psychological-safety interventions whether or not they are designed as such, which is the premise the rest of this paper builds on.

2.2 Algorithmic Management and the Machinery of Continuous Oversight

Kellogg, Valentine and Christin (2020) synthesise the algorithmic-management literature around six control mechanisms — restricting, recommending, recording, rating, replacing, and rewarding — through which algorithmic systems direct and evaluate labour, and they identify recording (continuous behavioural capture) as the mechanism from which most of the others draw their authority: a system can only recommend, rate, or reward on the basis of what it has first recorded. Wellbeing analytics sit squarely inside the recording mechanism, generalised from task performance to affective and physiological state, which is precisely the extension that makes existing algorithmic-management theory, developed mostly around productivity monitoring, only partially transferable to the wellbeing case.

Ravid, Tomczak, White and Behrend (2020), in a comprehensive review and research agenda for electronic performance monitoring, organise the field's accumulated findings around dimensions including monitoring intensity, the specificity of what is captured, and the degree of employee control over the monitoring process, and they conclude that these design dimensions — not the mere presence or absence of monitoring — are what determine whether employees experience a system as supportive oversight or invasive control. Alge (2001), in an earlier experimental test that anticipated much of this design-sensitivity, finds that computer surveillance perceived as capturing job-irrelevant information produces sharper reductions in perceived privacy and procedural justice than surveillance confined to job-relevant information, holding the underlying technology constant — evidence directly relevant to wellbeing monitoring, whose entire premise is the capture of information (affect, stress, exhaustion) that sits outside conventional task-relevance boundaries. Sastry's (2026a) analysis of algorithmic-management transparency mandates develops a parallel argument for a different outcome variable, employee trust in automated decisions, finding that the presence of a disclosure mandate does not by itself repair trust unless employees also judge the disclosed information to be genuinely actionable — a distinction between formal transparency and functional transparency that this paper carries into its treatment of wellbeing-monitoring design (Sastry, 2026a).

A second point of continuity is with AI adoption capability rather than AI design specifically. Sastry and Mallya's (2025b) study of AI analytics in hiring documents that HR functions frequently adopt AI tools for their efficiency case — faster, more consistent



processing of applicant or employee information — without a correspondingly developed capability to manage the relational consequences of that automation for the people being processed (Sastry & Mallya, 2025b). Wellbeing-monitoring tools are typically adopted through the same efficiency logic — continuous, low-marginal-cost signal at a scale no human observer could sustain — which suggests the relational blind spot Sastry and Mallya document for hiring analytics is likely to recur, and possibly to matter more, where the analytics concern an employee's psychological state rather than their qualifications.

2.3 Privacy, Trust and the Boundary of Legitimate Monitoring

Bhave, Teo and Dalal (2020), synthesising organisational-privacy research into a unifying framework, argue that workplace privacy is best understood not as a fixed entitlement but as a boundary employees actively negotiate, and that violations of the boundary — rather than monitoring as such — are what predict negative employee reactions; the same monitoring intensity can be experienced as intrusive or unremarkable depending on whether it falls inside or outside the boundary an employee believed was in place. This boundary-regulation framing is a natural fit for wellbeing monitoring, where the boundary in question is not merely how much is observed but what category of information — task output versus affective state — an employee reasonably expected the employer to have standing to observe at all.

Sewell and Barker (2006) supply the paper's central interpretive lens for why wellbeing monitoring is a harder case than productivity monitoring. Analysing organisational surveillance through the concept of situational irony, they show that the same monitoring practice can be sincerely intended by management as an act of care — protecting employees, supporting them, keeping them safe — while being experienced by the monitored as coercive control, with neither party's reading straightforwardly wrong; the irony is structural, arising from the asymmetry between who designs the monitoring and who is its object, not from bad faith on either side. Wellbeing monitoring is, on this account, the limiting case of the coercion-versus-care irony, because its stated intent (care) is more sincerely and more prominently foregrounded than in productivity-monitoring contexts, which if anything raises the stakes of misalignment between intended and received meaning rather than resolving it. Penney (2016), studying online surveillance disclosures and subsequent behaviour change, provides the empirical anchor for a further mechanism relevant here — a demonstrated chilling effect in which awareness of being observed measurably suppresses behaviour the observed party believes might be read unfavourably, evidence for exactly the suppression channel through which this paper argues wellbeing monitoring could depress voice even without any single punitive incident ever occurring.

The author's own work on AI-readiness practice bears on how this boundary is set inside an organisation rather than merely how it is experienced by individuals. Sastry (2026b) finds that the realised productivity return from agentic AI deployments depends less on the technology itself than on organisational readiness practices — workflow redesign, manager preparation, and explicit governance of what the system is and is not authorised to act on — and that firms deploying capable systems inside weak governance structures capture markedly less value from the deployment (Sastry, 2026b). Sastry and Mallya's (2025a) study of leadership strategies for young-talent management similarly finds that younger employees' trust in organisational practices is comparatively more contingent on the transparency and consistency of the practice than on its stated intent, which suggests that the goodwill built into a wellbeing-monitoring system's design brief travels only partway into how it is trusted once deployed (Sastry & Mallya, 2025a).

2.4 Voice, Silence, and the Costs of Speaking Up

Van Dyne and LePine (1998), in the study that established voice as a distinct extra-role construct from helping behaviour, define voice as the deliberate, promotive expression of ideas or concerns aimed at improving rather than merely criticising organisational functioning, and demonstrate that it is best predicted by employees' judgment of interpersonal risk rather than by their level of job satisfaction — an early empirical anchor for the risk-calculus mechanism this paper's model assumes. Morrison (2011), synthesising the subsequent voice-behaviour literature, confirms that perceived safety consistently outperforms other antecedents — including perceived efficacy of voicing — as a predictor of whether employees actually speak up, and specifically flags monitoring and surveillance practices as an under-examined antecedent class relative to leadership style and climate, a gap this paper addresses directly.

Newman, Fast and Harmon (2020) provide a mechanism for why algorithmically mediated organisational decisions can suppress voice even when they are procedurally more consistent than the human judgment they replace: employees judge algorithmic decision processes as less fair when those processes are perceived to strip out the contextual, individuating information a human decision-maker would have used, an effect the authors term algorithmic reductionism. Wellbeing-monitoring outputs are a plausible site for this same reductionism — a burnout-risk score compresses an employee's full situation into a single number — and if employees anticipate that raising a concern will be filtered through, or second-guessed against, that reductive score, the incentive to raise the concern through the ordinary judgment-rich channel of a manager conversation is correspondingly weakened.

The author's work on delegation and employee initiative offers a complementary voice mechanism grounded not in fear of the monitoring system but in the erosion of felt ownership over the problem-raising process. Sastry and Ravish (2025a), studying delegation and resource provision in the Indian workplace, find that employees are more willing to raise operational concerns when they retain some agency over how the concern is acted on rather than having a fixed remediation pathway imposed on them (Sastry & Ravish, 2025a); wellbeing-monitoring systems that auto-route a flagged employee into a standardised support pathway can inadvertently reproduce exactly this loss of agency. Sastry and Ravish (2025b), extending the analysis to decision-making authority



more broadly, similarly find that employee initiative is measurably higher where employees perceive that raising an issue changes what happens next in a way they can see, rather than disappearing into an opaque process — a visibility condition that an algorithmically mediated wellbeing-escalation pathway can either satisfy or violate depending on its design (Sastry & Ravish, 2025b).

2.5 Wellbeing Analytics and the Paradox of Caring Surveillance

The empirical literature specifically on AI-mediated wellbeing and burnout-prediction tools is younger and thinner than the adjacent literatures reviewed above, but a consistent finding emerges from the closest available evidence on continuous workplace monitoring and worker wellbeing more generally: Glavin, Bierman and Schieman (2024), examining the relationship between the extent of employer monitoring practices and worker wellbeing outcomes, document that more intensively monitored workers report systematically poorer wellbeing indicators than less intensively monitored workers holding job characteristics constant, a finding that sits in direct tension with the productivity-monitoring literature's usual justification that monitoring exists to protect and support the workforce it is applied to. Tambe, Cappelli and Yakubovich (2019), reviewing the broader promise and limits of AI in human resource management, caution that HR analytics applications are frequently deployed ahead of the validation evidence that would justify confidence in their predictive accuracy, a caution with particular force for burnout-prediction models, whose false positives and false negatives both carry a psychological-safety cost — the former by subjecting an unaffected employee to unwanted scrutiny, the latter by allowing the organisation's stated commitment to “catching problems early” to go visibly unfulfilled for an employee who was in fact struggling.

This is the paradox the paper's title names: caring surveillance is not a contradiction in terms from the deploying organisation's perspective, but it is functionally continuous, from the monitored employee's perspective, with the surveillance apparatus Sewell and Barker (2006) and Penney (2016) describe, and the same design features — continuous capture, individual attribution, opaque decision logic — that make a monitoring system operationally effective at identifying risk are the features that make it experientially indistinguishable from coercive oversight. The paper's proposition set in Section 4 is built around the conditions under which this indistinguishability can be reduced without abandoning the monitoring capability altogether.

The author's own work on leadership and burnout, and on social comparison inside organisations, both bear directly on this paradox. Sastry (2024), examining leadership practices aimed at combating workforce burnout, finds that interventions perceived by employees as genuinely developmental — oriented toward building the employee's own capacity rather than toward organisational risk management — produce measurably better engagement outcomes than interventions with identical content but a compliance-oriented framing (Sastry, 2024), a framing-sensitivity result this paper treats as directly transferable to how a burnout-prediction alert is communicated once generated. Sastry (2025), studying talent designation and social comparison, finds that being singled out by an organisational system — even for ostensibly positive reasons, such as identification as high-potential — changes how an employee is subsequently perceived by peers and how the employee subsequently manages their own visibility, a mechanism that applies with equal or greater force to being singled out by a burnout-risk flag, where the social comparison stakes run in the opposite, stigmatising direction (Sastry, 2025).

2.6 The India IT-BPM Context

India's information-technology and business-process-management sector is a distinctive setting for wellbeing-monitoring adoption for three joint reasons: the sector's productivity- and utilisation-tracking infrastructure is already extensive, meaning wellbeing analytics are typically layered onto an existing monitoring stack rather than introduced as a standalone system; attrition has historically run high enough that employee wellbeing carries a direct, board-visible cost narrative that gives wellbeing-analytics adoption an unusually strong internal business case; and the sector's shift-based, often 24x7 delivery model against international time zones creates exactly the workload-irregularity signal — unusual hours, compressed turnaround, after-hours communication spikes — that burnout-prediction models are built to detect, making the sector both an early and a heavily instrumented adopter.

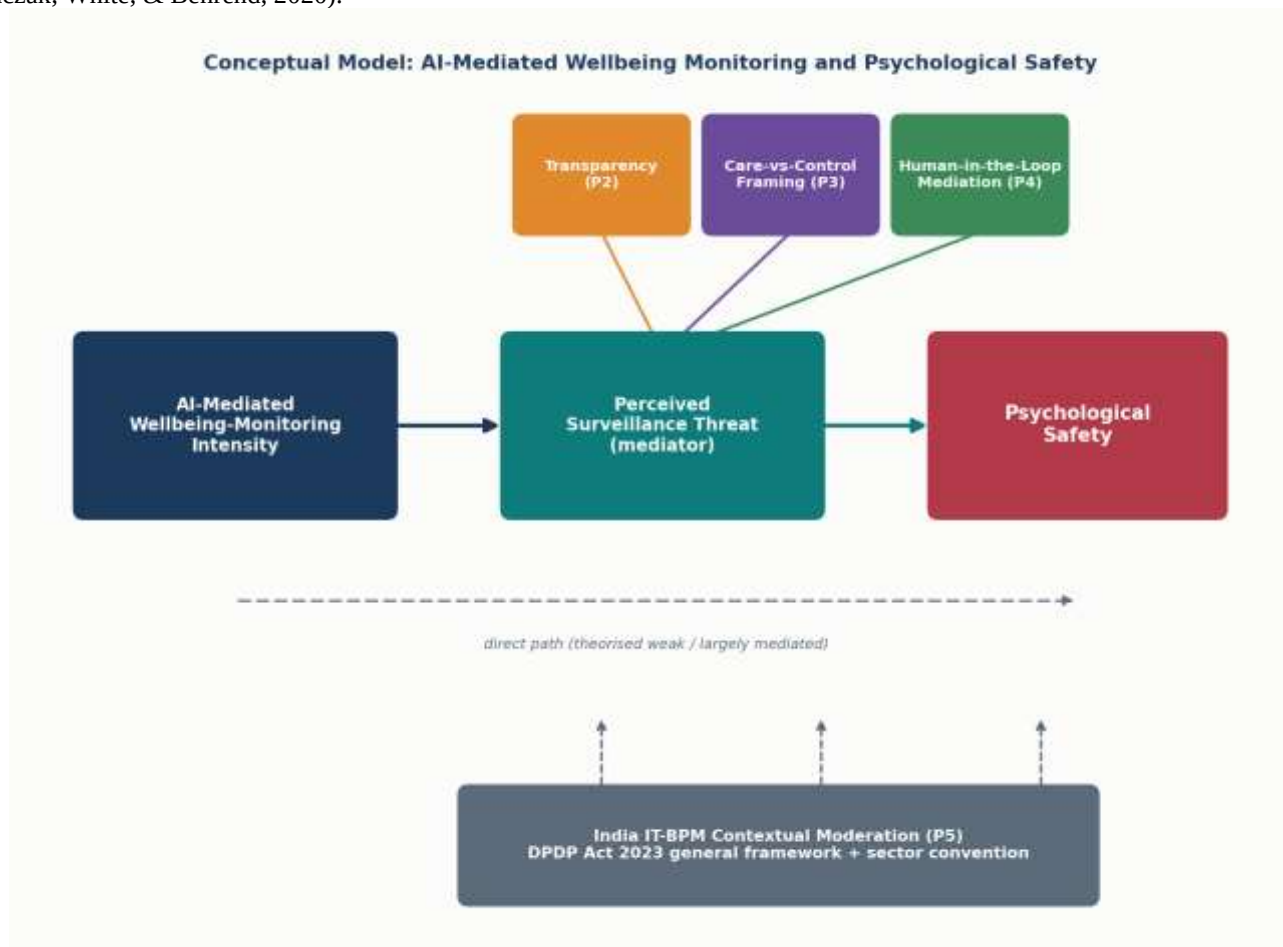
The regulatory backdrop is the Digital Personal Data Protection Act, 2023 (DPDP Act), which establishes a general framework of consent, purpose limitation, and data minimisation applicable to any processing of personal data, employee wellbeing data included, but does not contain a provision specific to workplace wellbeing analytics, sentiment monitoring, or burnout-prediction scoring. Employers processing this category of data are required, as under the general framework, to obtain informed consent, disclose the purpose of collection, limit processing to what the stated purpose requires, and maintain accountability records including impact assessments for higher-risk processing activities — obligations that apply with full force to wellbeing-monitoring systems even though no wellbeing-specific rule sits on top of them. This produces a compliance environment structurally similar to the one this series' companion paper on pay-transparency mandates identified for compensation disclosure: a general legal backdrop shapes the outer boundary of what is permitted, while the specific norms governing a particular HR practice are set by employer and sector convention rather than by statute.

The author's work on risk modelling in dynamic environments and on performance-appraisal engagement both speak to how this convention-setting process is likely to unfold. Sastry and Arunachalam (2025), developing a probabilistic framework for managing risk under uncertainty in AI-governed systems, argue that risk-scoring systems deployed without an explicit, disclosed tolerance for false positives tend to be calibrated conservatively by the organisations that operate them — flagging more rather than fewer borderline cases — precisely because the reputational cost of a missed case is more salient to the organisation than the psychological cost of an unnecessary flag to the individual (Sastry & Arunachalam, 2025), a calibration asymmetry with direct relevance to how

liberally an India IT-BPM burnout-prediction system is likely to be tuned. Sastry and Selvarasu (2014), in an earlier study of performance appraisal and employee engagement, find that engagement is measurably more sensitive to whether an evaluative system is perceived as procedurally fair than to the favourability of the specific rating it produces (Sastry & Selvarasu, 2014), a finding this paper reads as evidence that the procedural design of a wellbeing-monitoring escalation pathway — not merely its intended benevolence — will determine its reception inside India's IT-BPM workforce.

3. CONCEPTUAL FRAMEWORK

Figure 1 sets out the paper's core causal structure. AI-mediated wellbeing-monitoring intensity — the breadth, granularity, and continuity of what a monitoring system observes about an employee's affective and behavioural state — is proposed to act on psychological safety principally through a mediating perception of surveillance threat: the employee's belief that what the system observes about them could be used against their interests. This mediated path is consistent with the psychological-safety literature's own emphasis on shared belief rather than objective practice as the proximal driver of voice behaviour (Edmondson, 1999; Edmondson & Lei, 2014), and with the surveillance literature's demonstration that the same underlying monitoring technology produces divergent employee reactions depending on how its purpose and reach are perceived (Bhave, Teo, & Dalal, 2020; Ravid, Tomczak, White, & Behrend, 2020).



Three variables are proposed to moderate the strength of the monitoring-intensity to surveillance-threat link: the transparency of the monitoring system (how clearly employees can see what is captured, by whom, and for what stated purpose), the developmental-versus-punitive framing of the system's use in practice (whether flagged outputs visibly lead to support or to consequence), and the presence of human mediation between algorithmic output and managerial action (whether a manager interprets and contextualises a flag before acting, or the system triggers action directly). A fourth relationship specifies a contextual boundary condition for India's IT-BPM sector, where regulatory generality and sector-specific labour-market pressure combine to shape how the first three moderators are likely to be resolved in practice. Table 1 defines the framework's key constructs.



Table 1. Key Construct Definitions

Construct	Working Definition (as used in this paper)	Representative Source
Psychological safety	A shared team-level belief that the environment is safe for interpersonal risk-taking, including voicing dissent, admitting error, or raising concerns.	Edmondson (1999)
AI-mediated wellbeing monitoring	Continuous or near-continuous algorithmic capture and interpretation of employee communication, workload, or behavioural data to infer affective or physiological state.	Ravid, Tomczak, White, & Behrend (2020)
Sentiment analytics	Natural-language-processing tools that classify the emotional tone of written workplace communication (e.g., email, chat).	Kellogg, Valentine, & Christin (2020)
Burnout-prediction tools	Models that combine behavioural and workload signals into an individual risk score for exhaustion or disengagement.	Tambe, Cappelli, & Yakubovich (2019)
Perceived surveillance threat	An employee's belief that information observed about them could be used in ways contrary to their interests.	Bhave, Teo, & Dalal (2020)
Coercion–care irony	The structural condition in which a monitoring practice is sincerely intended as care by management while being experienced as control by the monitored.	Sewell & Barker (2006)

4. PROPOSITIONS

The propositions below move from the framework's core mediated relationship (P1) through three design-sensitive moderators (P2–P4) to a contextual boundary condition (P5). Each proposition is stated at the level of the organisational practice rather than the individual employee, consistent with psychological safety's status as a collective, climate-level construct (Edmondson, 1999). Table 2 first compares the principal wellbeing-monitoring modalities in current use against the design dimensions the propositions draw on, since P2–P4 apply differently depending on which modality, or combination of modalities, an organisation has deployed.

Table 2. Wellbeing-Monitoring Modalities Compared

Modality	Data Granularity	Individual vs. Aggregate	Typical Human-in-the-Loop	Typical Disclosure Level
Communication sentiment analysis	High (message-level)	Individual, aggregated for reporting	Low – alerts often route directly to HR/manager dashboards	Variable – often disclosed as policy, rarely as message-level scope
Calendar / workload digital-exhaust analytics	Medium (activity-pattern level)	Individual and team	Medium – usually reviewed before escalation	Moderate – workload metrics commonly named in policy
Wearable / biometric-adjacent signals	Very high (continuous physiological proxy)	Individual	Low to medium, technology-dependent	Low – scope of inference often underspecified to users
Aggregated pulse-survey plus AI synthesis	Low (periodic, self-reported)	Primarily aggregate	High – typically reviewed by HR before action	High – survey purpose usually explicit
Composite burnout-risk scoring	High (multi-signal composite)	Individual	Variable – the design dimension this paper's P4 addresses	Variable – the design dimension this paper's P2 addresses

Proposition 1 (core mediation). The intensity of AI-mediated wellbeing monitoring is negatively associated with team-level psychological safety, and this relationship is mediated by employees' perceived surveillance threat rather than operating as a direct effect of monitoring intensity itself. This follows from the theoretical logic developed in Sections 2.1 and 2.3: monitoring intensity has no fixed psychological meaning independent of how it is interpreted (Bhave, Teo, & Dalal, 2020), and it is the interpretation — codified here as perceived surveillance threat — that supplies the mechanism through which Detert and Edmondson's (2011) implicit voice theories are updated. A purely direct-effect specification would imply that psychological safety responds mechanically to monitoring intensity regardless of how the monitoring is understood, which the boundary-regulation and situational-irony literatures reviewed above (Bhave, Teo, & Dalal, 2020; Sewell & Barker, 2006) suggest is inconsistent with how employees actually process organisational surveillance.

Proposition 2 (transparency moderation). The negative association between monitoring intensity and psychological safety is attenuated where the monitoring system is high in functional transparency — where employees can accurately state what is captured, who can see it, and what triggers an escalation — relative to where the system is procedurally opaque even if formally disclosed. This distinction between formal and functional transparency follows directly from Sastry's (2026a) finding that disclosure mandates alone do not repair trust in automated decisions unless the disclosed information is genuinely actionable; a wellbeing-monitoring system can satisfy a disclosure requirement in form while remaining functionally opaque to the employees it monitors, in which case P2's attenuating effect should not be observed.

Proposition 3 (care-versus-control framing moderation). The negative association between monitoring intensity and psychological safety is attenuated where employees attribute developmental rather than punitive intent to the monitoring system's use in practice, and this attribution is shaped more strongly by observed downstream consequences of being flagged than by the system's stated design purpose. This proposition operationalises the Sewell and Barker (2006) coercion-versus-care irony as an empirically tractable moderator: the same stated purpose (care) can resolve toward either pole depending on what employees actually observe happening to colleagues who are flagged, consistent with Sastry's (2024) finding that developmentally framed interventions with identical content to compliance-framed interventions produce systematically different engagement outcomes.



Proposition 4 (human-in-the-loop moderation). The negative association between monitoring intensity and psychological safety is attenuated where a human manager interprets and contextualises algorithmic wellbeing outputs before any action is taken, relative to architectures in which flagged outputs trigger automated or procedurally fixed action. This follows from Newman, Fast and Harmon's (2020) algorithmic-reductionism mechanism: human mediation restores exactly the contextual, individuating information whose absence employees penalise when judging algorithmic decisions as unfair, and it is consistent with Sastry and Ravish's (2025b) finding that employee initiative is higher where the path from raising a concern to organisational action remains visible and non-mechanical rather than disappearing into an opaque, fixed pathway.

Proposition 5 (India IT-BPM contextual moderation). In India's IT-BPM sector, the attenuating effects specified in P2–P4 are weaker on average than in settings with a sector-specific wellbeing-monitoring mandate, because the DPDP Act's general consent-and-purpose-limitation framework leaves the specific transparency, framing, and human-mediation practices governing wellbeing analytics to be set by employer convention rather than by binding sector rule; consequently, the erosion of psychological safety associated with high monitoring intensity is expected to be most pronounced among IT-BPM employers that have layered wellbeing analytics onto an existing productivity-monitoring stack without independently redesigning transparency, framing, or escalation practice for the wellbeing use case specifically. This is



consistent with Sastry and Arunachalam's (2025) finding that risk-scoring systems tend to be calibrated conservatively — toward more rather than fewer flags — in the absence of an explicit, disclosed tolerance for false positives, and with Sastry and Selvarasu's (2014) finding that engagement in Indian organisational settings tracks procedural fairness more closely than the favourability of any single evaluative output.

5. ILLUSTRATIVE EVIDENCE

No published study yet tests this paper's full mediated-moderation model directly, which is expected given how recently AI-mediated wellbeing tools have moved from pilot to routine deployment; the purpose of this section is instead to show that the framework's separate components are each consistent with the closest available evidence, which is the standard this paper's companion pieces in the series have applied to conceptual, propositions-based work (Sastry, 2026b). Table 3 summarises five bodies of evidence directly relevant to the model.

Table 3. Illustrative Evidence Consistent with the Framework

Source	Context	Finding Relevant to This Paper's Model
Ravid, Tomczak, White, & Behrend (2020)	Review and research agenda, electronic performance monitoring	Design dimensions (intensity, specificity, employee control) — not monitoring's mere presence — predict whether monitoring is experienced as supportive or invasive.
Glavin, Bierman, & Schieman (2024)	Survey study, workplace surveillance and worker wellbeing	More intensively monitored workers report systematically poorer wellbeing indicators, holding job characteristics constant.
Alge (2001)	Experimental study, computer surveillance	Surveillance perceived as capturing job-irrelevant information produces sharper reductions in perceived privacy and procedural justice than job-relevant surveillance, holding technology constant.
Bhave, Teo, & Dalal (2020)	Integrative review, workplace privacy	Privacy violations — not monitoring intensity alone — predict negative employee reactions, framing privacy as a negotiated boundary.
Newman, Fast, & Harmon (2020)	Experimental study, algorithmic HR decisions	Employees judge algorithmic decisions as less fair when perceived to strip out individuating context a human decision-maker would have used (algorithmic reductionism).

Three patterns in this evidence base are worth drawing out explicitly. First, the consistent finding across Ravid, Tomczak, White and Behrend (2020) and Glavin, Bierman and Schieman (2024) that monitoring's design dimensions — not its bare presence — predict employee reaction supports this paper's decision to model transparency and framing as moderators of a mediated relationship rather than as direct effects competing with monitoring intensity for explanatory weight. Second, Alge's (2001) finding that job-relevance perceptions drive reactions to surveillance more than the surveillance technology itself does is a close structural analogue to this paper's prediction that developmental-versus-punitive framing (P3), not monitoring intensity in isolation, will carry much of the explanatory weight for psychological-safety outcomes. Third, Newman, Fast and Harmon's (2020) algorithmic-reductionism finding, generated in a hiring-decision context, transfers to the wellbeing-monitoring context studied here only if burnout-risk scores are in fact experienced by employees as reductive summaries of a fuller situation — an assumption the paper treats as plausible given the scoring logic such tools typically use, but one that future empirical work, discussed in Section 9, should test directly rather than assume.



Figure 3 summarises this evidentiary pattern as a care-control quadrant defined by monitoring transparency and perceived intent. The upper-right quadrant — high transparency, developmentally framed — is the region in which the literature reviewed above would predict monitoring and psychological safety can coexist; the lower-left quadrant — opaque, punitively framed — is the region structurally closest to the coercive surveillance Sewell and Barker (2006) and Penney (2016) describe. The two off-diagonal quadrants are analytically interesting because they represent partial, unstable configurations — a transparent but punitively resolved system, or an opaque but sincerely developmental one — that this paper's model predicts will drift toward one of the diagonal poles as employees update their implicit voice theories from observed rather than stated practice (Detert & Edmondson, 2011).

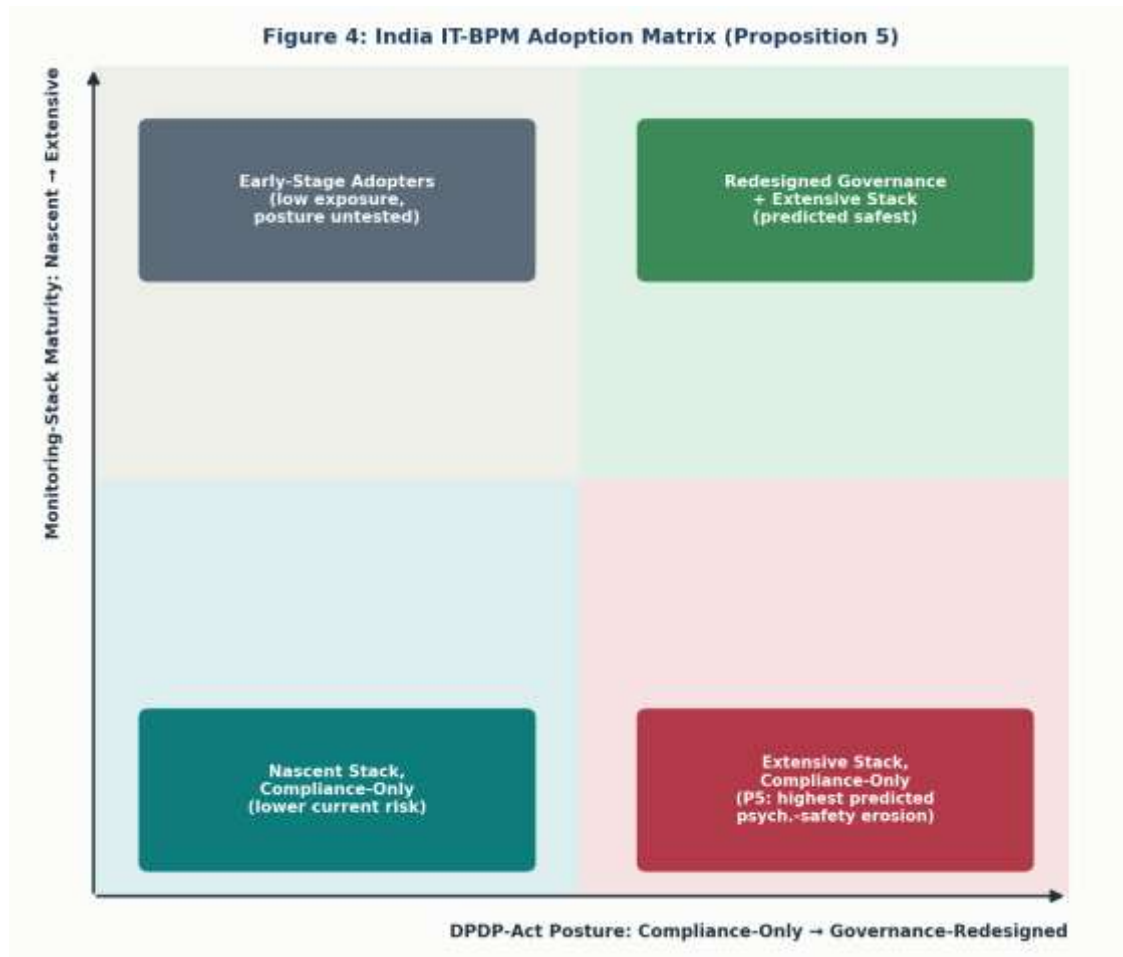
6. APPLICATION TO INDIA'S IT-BPM SECTOR

Section 2.6 established the regulatory and labour-market backdrop against which India's IT-BPM employers are adopting wellbeing analytics; this section develops the application in more operational detail, focused on how Proposition 5's contextual moderation is likely to manifest inside the sector's existing organisational practice. Table 4 summarises the comparison between a generic global adoption pattern and the pattern this paper expects to be more common in India's IT-BPM sector specifically.

Table 4. Global vs. India IT-BPM Adoption Pattern

Dimension	Generic Global Pattern	India IT-BPM Pattern (this paper's expectation)
Monitoring infrastructure	Wellbeing analytics often introduced as a standalone initiative	Layered onto an already-extensive productivity/utilisation-tracking stack
Regulatory backdrop	Varies widely; some jurisdictions have workplace-monitoring-specific rules	DPDP Act 2023 — general consent/purpose-limitation framework, no wellbeing-monitoring-specific rule
Adoption driver	Mixed — productivity, risk management, and wellbeing rationales	Strong attrition-cost narrative gives wellbeing-analytics adoption an unusually direct business case

Workload signal source	Varies by sector	Shift-based, 24x7, cross-time-zone delivery model generates strong, continuous workload-irregularity signals
Governance maturity (P5 driver)	Uneven, sector- and firm-dependent	Expected to bifurcate: DPDP compliance-only postures vs. employers using impact assessments to redesign wellbeing-specific governance



The practical implication of Figure 4's matrix is that the same wellbeing-monitoring technology can land in structurally different quadrants across employers within a single sector, depending on whether the employer treats DPDP-Act compliance as a floor to build psychologically safe practice on top of, or as the ceiling of what wellbeing-monitoring governance requires. Employers with mature productivity-monitoring infrastructure and a compliance-only posture toward the DPDP Act are, on this paper's model, the population in which P5's predicted erosion of psychological safety should be most visible, because they possess the monitoring intensity the model requires without having independently built the transparency, framing, and human-mediation practices P2–P4 identify as attenuating. Employers that have used the DPDP Act's impact-assessment requirement as an occasion to redesign wellbeing-monitoring governance specifically — rather than treating it as a documentation exercise — are the population in which the model predicts monitoring intensity and psychological safety are more likely to coexist.

7. SYNTHESIS

Figure 5 consolidates the paper's five propositions into a single summary map, showing the core mediated relationship (P1) running through perceived surveillance threat to psychological safety, the three design-sensitive moderators (P2–P4) acting on the strength of that mediated path, and the India IT-BPM contextual moderation (P5) acting on the average strength of P2–P4 themselves rather than on P1 directly — a second-order moderation structure that follows from treating regulatory and sectoral context as shaping which design choices organisations actually make, rather than as an independent driver of psychological safety in its own right.

Figure 5: Propositions Summary Map

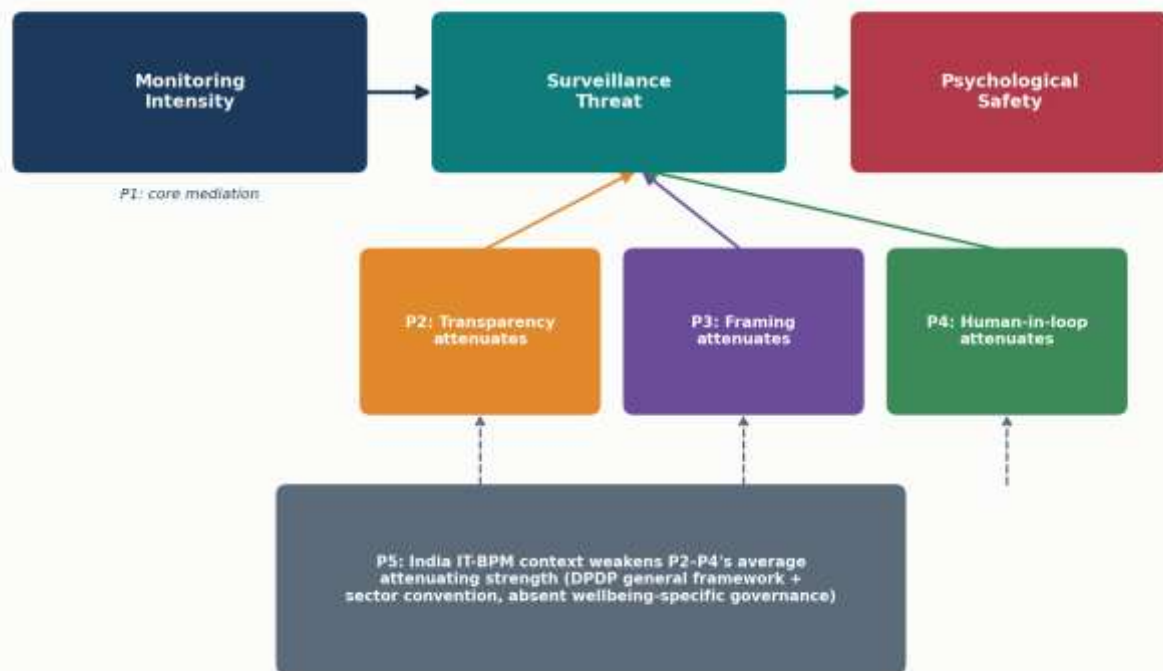


Table 5 restates the five propositions in summary form alongside their proposed mechanism, moderator, and expected direction, to serve as a single reference point for the empirical work outlined in Section 9.

Table 5. Summary of Propositions

#	Statement (summary)	Mechanism	Moderator / Condition	Expected Direction
P1	Monitoring intensity reduces psychological safety via perceived surveillance threat	Mediation	— (core path)	Negative, mediated
P2	Effect is weaker where the system is functionally (not just formally) transparent	Attribution / sense-making	Functional transparency	Attenuating
P3	Effect is weaker where employees observe developmental rather than punitive consequences of being flagged	Framing / observed practice	Care-vs-control framing	Attenuating
P4	Effect is weaker where a human manager mediates between algorithmic output and action	Restored contextual judgment	Human-in-the-loop design	Attenuating
P5	Attenuating effects of P2–P4 are weaker on average in India IT-BPM absent deliberate wellbeing-specific governance	Regulatory generality + sector convention	India IT-BPM context	Second-order moderation

8. DISCUSSION AND THEORETICAL IMPLICATIONS

This paper's central theoretical contribution is to specify psychological safety as the construct through which AI-mediated wellbeing monitoring's caring intent and surveillance structure are reconciled or fail to be reconciled, rather than treating the two framings as competing descriptions of the same practice. The algorithmic-management literature has developed rich accounts of how monitoring shapes task behaviour (Kellogg, Valentine, & Christin, 2020) and the workplace-privacy literature has developed rich accounts of when monitoring is experienced as a boundary violation (Bhave, Teo, & Dalal, 2020), but neither literature, on its own, specifies what continuous affective and behavioural monitoring does to the specific organisational capacity — voice — that most depends on employees' trust that the organisation is safe to be candid with. Connecting these literatures through psychological safety, rather than through job satisfaction or turnover intention as the



outcome of interest, foregrounds a cost of poorly designed wellbeing monitoring that is easy for organisations to miss precisely because it does not show up as a complaint: silence, definitionally, does not announce itself.

A second theoretical implication concerns the paper's second-order moderation structure in Proposition 5. Most conceptual work applying Western-developed organisational-behaviour constructs to the Indian context treats national or sectoral context as a first-order moderator acting directly on the focal relationship. This paper's model instead treats the India IT-BPM context as shaping which values the first-order moderators (transparency, framing, human mediation) are likely to take in practice, which is a different and, this paper argues, more portable claim: it does not assert that Indian employees respond differently in principle to transparent versus opaque monitoring, but that the regulatory and competitive environment they work within makes transparent, developmentally framed, human-mediated monitoring systems less likely to be the ones actually deployed. This distinction matters for how the proposition should be tested — as a claim about the distribution of organisational practice, not about a cross-cultural difference in underlying psychological response.

Third, the paper's treatment of algorithmic reductionism (Newman, Fast, & Harmon, 2020) as transferable from hiring decisions to wellbeing-monitoring outputs is a theoretical extension worth flagging as provisional. Hiring decisions and burnout-risk flags differ in an important respect: a hiring decision is a single, discrete, high-stakes event, while a burnout-risk flag is typically one signal inside an ongoing, revisable relationship between an employee and their manager. Whether algorithmic reductionism operates with the same force in a revisable, relational context as in a discrete, terminal one is an open question the framework raises rather than resolves, and Section 9 identifies it as a priority for empirical work.

9. MANAGERIAL AND POLICY IMPLICATIONS

For organisations already operating or actively evaluating AI-mediated wellbeing-monitoring tools, this paper's framework suggests three concrete design priorities, corresponding to Propositions 2 through 4. The transparency priority is to close the gap between formal and functional transparency identified in Section 2.3: publishing what a system captures is not equivalent to employees being able to accurately state what it captures, and organisations serious about the transparency moderator should test employee comprehension of the monitoring system directly rather than treating a disclosure document's existence as sufficient. The framing priority is to recognise, following Sewell and Barker (2006) and Sastry (2024), that a system's stated developmental purpose is confirmed or disconfirmed for employees by what they observe happening to flagged colleagues, which means the single highest-leverage design choice available to an organisation is what the first several flagged cases are visibly seen to result in. The human-mediation priority is to preserve a human interpretive step between algorithmic output and managerial action even where full automation is technically available, on the strength of the algorithmic-reductionism evidence reviewed in Section 2.4 and Section 5.

For HR and people-analytics functions specifically, the paper's synthesis of Sastry and Mallya's (2025b) finding on the relational blind spot in AI-adoption practice and Sastry's (2026b) finding on AI-readiness governance suggests that wellbeing-monitoring tools should not be evaluated by the same efficiency metrics — coverage, detection rate, processing cost — used for productivity-monitoring or hiring-analytics tools, because those metrics are silent on the psychological-safety cost the framework identifies. A minimal governance addition consistent with this paper's model is to track, alongside detection-rate metrics, whether voice and error-reporting rates change following wellbeing-monitoring deployment — a directly observable proxy for the psychological-safety outcome the model predicts monitoring can either protect or erode.

For India IT-BPM policymakers and sector bodies, the paper's account of the DPDP Act's general-but-not-sector-specific coverage suggests that the Act's existing impact-assessment requirement for high-risk processing could be extended, through sector guidance rather than new primary legislation, to require that a wellbeing-monitoring impact assessment specifically evaluate transparency, framing, and human-mediation design — the three levers this paper's model identifies as attenuating psychological-safety erosion — rather than treating the assessment as satisfied by a generic data-protection review. This would convert Proposition 5's predicted vulnerability, that general compliance leaves wellbeing-specific design norms unset, into a specific, auditable governance requirement without mandating any particular monitoring technology or prohibiting the practice outright.

10. LIMITATIONS AND FUTURE RESEARCH

As a conceptual paper, this work's central limitation is that its propositions are derived from theoretical integration and adjacent empirical literatures rather than from direct measurement of the monitoring-intensity to psychological-safety relationship, which no published study has yet tested as specified here. Priority empirical work should include a validated measure of AI-mediated wellbeing-monitoring intensity distinct from general electronic-performance-monitoring scales, since existing EPM instruments were developed around productivity and task-completion monitoring and may not adequately capture the affective-inference dimension this paper's framework treats as central; a longitudinal or quasi-experimental design around the introduction of a wellbeing-monitoring system, to establish temporal precedence for the



proposed mediation rather than relying on cross-sectional association; and a direct test of whether algorithmic reductionism (Newman, Fast, & Harmon, 2020) operates with comparable force in the revisable, relational context of an ongoing wellbeing flag as in the discrete hiring-decision context in which it was originally documented, the open question flagged in Section 8.

A second limitation concerns scope: the paper's framework is developed for organisational-level wellbeing-monitoring tools operating on communication and workload data, and does not extend without further theoretical work to monitoring modalities involving biometric or physiological sensing, which raise a distinct set of consent and bodily-autonomy considerations beyond this paper's scope. A third limitation is that Proposition 5's India IT-BPM application is developed from regulatory analysis and the author's prior empirical work on related but not identical constructs (delegation, risk-scoring calibration, appraisal fairness) rather than from primary data on wellbeing-monitoring practice inside the sector specifically; a sector survey directly measuring monitoring intensity, transparency, framing, and human-mediation practice across a sample of India IT-BPM employers, paired with team-level psychological-safety measures, is the natural next step in this paper's research programme and a candidate for the empirical companion study this conceptual paper is intended to motivate.

11. CONCLUSION

AI-mediated wellbeing monitoring is neither the unambiguous organisational good its adoption rationale typically claims nor the straightforward surveillance threat a purely critical reading would suggest; this paper's argument is that which of those two descriptions comes to govern an employee's lived experience of being monitored is a design outcome, not a fixed property of the technology, and that psychological safety is the construct through which that design outcome is felt organisationally. The five propositions developed here specify transparency, developmental framing, and human mediation as the levers through which organisations can move a wellbeing-monitoring system toward the caring pole of Sewell and Barker's (2006) coercion-versus-care irony rather than the coercive one, and specify India's IT-BPM sector as a setting in which general data-protection compliance, absent deliberate attention to those three levers, is unlikely to be sufficient to secure that outcome on its own.

The practical stakes of getting this design right are not confined to employee wellbeing narrowly construed. Because psychological safety is the mechanism through which organisations learn about the problems inside them — the errors, the near-misses, the emerging risks employees are best positioned to see first — a wellbeing-monitoring architecture that inadvertently narrows the space employees judge safe for candour imposes an organisational learning cost alongside any individual wellbeing cost. Organisations adopting these tools with the stated aim of catching problems early have, on this paper's argument, a direct interest in the design choices this paper identifies, independent of any regulatory obligation to make them.

Note on Manuscript: This paper is offered as a propositions-based conceptual contribution in the format the author has used across this HRM-variable series, integrating psychological-safety theory, the algorithmic-management and workplace-surveillance literatures, and organisational-voice research around a practice — AI-mediated wellbeing monitoring — that has moved from pilot to routine deployment faster than the theory needed to evaluate its organisational consequences. All external sources cited were verified against publisher, aggregator, or crossref bibliographic records at the time of writing; where an exact statistic could not be independently confirmed, the paper cites the source for its documented scope or general finding only, consistent with the accuracy standard applied across this series. Empirical validation of the propositions advanced here, particularly the mediation specified in Proposition 1 and the India IT-BPM application specified in Proposition 5, is identified in Section 9 as the natural next stage of this research programme.

REFERENCES

1. Alge, B. J. (2001). *Effects of computer surveillance on perceptions of privacy and procedural justice*. *Journal of Applied Psychology*, 86(4), 797–804.
2. Bhawe, D. P., Teo, L. H., & Dalal, R. S. (2020). *Privacy at work: A review and a research agenda for a contested terrain*. *Journal of Management*, 46(1), 127–164.
3. Detert, J. R., & Edmondson, A. C. (2011). *Implicit voice theories: Taken-for-granted rules of self-censorship at work*. *Academy of Management Journal*, 54, 461–488.
4. Edmondson, A. C. (1999). *Psychological safety and learning behavior in work teams*. *Administrative Science Quarterly*, 44(2), 350–383.
5. Edmondson, A. C., & Lei, Z. (2014). *Psychological safety: The history, renaissance, and future of an interpersonal construct*. *Annual Review of Organizational Psychology and Organizational Behavior*, 1, 23–43.
6. Glavin, P., Bierman, A., & Schieman, S. (2024). *Private eyes, they see your every move: Workplace surveillance and worker wellbeing*. *Social Currents*, 11(4), 327–345.
7. 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.



8. Morrison, E. W. (2011). *Employee voice behavior: Integration and directions for future research*. *Academy of Management Annals*, 5(1), 373–412.
9. Morrison, E. W., & Milliken, F. J. (2000). *Organizational silence: A barrier to change and development in a pluralistic world*. *Academy of Management Review*, 25, 706–725.
10. Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). *When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions*. *Organizational Behavior and Human Decision Processes*, 160, 149–167.
11. Penney, J. W. (2016). *Chilling effects: Online surveillance and Wikipedia use*. *Berkeley Technology Law Journal*, 31(1), 117–182.
12. 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.
13. Sastry, N. S. K. (2024). *Revitalizing leadership to combat workforce burnout and improve performance*. *EPRA International Journal of Environmental Economics, Commerce and Educational Management (EPRA IJEECEM)*, 11(9).
14. Sastry, N. S. K. (2025). *Talent designation and social comparison: Economic drivers of organizational growth*. *EPRA International Journal of Economics, Business and Management Studies (EPRA IJEBMS)*, 12(10).
15. Sastry, N. S. K. (2026a). *Algorithmic management transparency mandates and employee trust in automated decisions*. Manuscript under review.
16. Sastry, N. S. K. (2026b). *Organisational AI-readiness practice and the realised productivity gap from agentic AI*. Manuscript under review.
17. Sastry, N. S. K., & Arunachalam, T. (2025). *Probabilistic safe RL: Modelling and managing risk in dynamic environments*. *EPRA International Journal of Research and Development (EPRA IJRD)*, 10(9), 83–87.
18. Sastry, N. S. K., & Mallya, M. M. (2025a). *Building future-ready organizations: Leadership strategies for young talent management*. *EPRA International Journal of Research and Development (EPRA IJRD)*, 10(9), 88–91.
19. Sastry, N. S. K., & Mallya, M. M. (2025b). *Revolutionizing HR hiring with AI analytics*. *EPRA International Journal of Research and Development (EPRA IJRD)*, 10(9), 125–131.
20. Sastry, N. S. K., & Ravish, P. Y. (2025a). *Empowering employees in the Indian workplace: The transformative power of delegation and resources*. *EPRA International Journal of Research and Development (EPRA IJRD)*, 10(7).
21. Sastry, N. S. K., & Ravish, P. Y. (2025b). *The impact of delegated authority and resource provision on decision-making – Fostering employee initiative*. *EPRA International Journal of Multidisciplinary Research (EPRA IJMR)*, 11(7).
22. Sastry, N. S. K., & Selvarasu, A. (2014). *A study of impact on performance appraisal on employee's engagement*. *International Journal of Managerial Studies and Research*, 2(11), 10–22.
23. Sewell, G., & Barker, J. R. (2006). *Coercion versus care: Using irony to make sense of organizational surveillance*. *Academy of Management Review*, 31(4), 934–961.
24. Tambe, P., Cappelli, P., & Yakubovich, V. (2019). *Artificial intelligence in human resources management: Challenges and a path forward*. *California Management Review*, 61(4), 15–42.
25. Van Dyne, L., & LePine, J. A. (1998). *Helping and voice extra-role behaviors: Evidence of construct and predictive validity*. *Academy of Management Journal*, 41(1), 108–119.