INTERNATIONAL JOURNAL OF MANAGEMENT AND ORGANIZATIONAL RESEARCH

The Algorithmic Empathy–Authenticity Gap: A Perspective on the Dark Side of AI-Enabled Empathy in Leadership

Anurag Tiruwa [1*], Shuchi Dikshit [2]
[1] Assistant Professor, Institute of Information Technology and Management, GGSIPU, Delhi, India
[2] Associate Professor, Jagan Institute of Management Studies, Delhi, India

* Corresponding Author: **Anurag Tiruwa**

Article Info

ISSN (online): 2583-6641
Impact Factor (RSIF): 8.56
Volume: 05
Issue: 04
Received: 23-04-2026
Accepted: 25-05-2026
Published: 27-06-2026
Page No: 09-14

Abstract
Artificial intelligence tools that detect, infer or simulate emotion are rapidly entering the leadership toolkit, promising earlier detection of burnout, personalised check-ins and scalable care. Yet when empathy is increasingly mediated by algorithms, core relational qualities of leadership—authenticity, trust and psychological safety—can be put at risk. This paper explores the dark side of AI-enabled empathy by conceptualising an algorithmic empathy–authenticity gap: a disconnect between the growing volume and precision of empathetic signals and employees' sense that those signals are genuinely felt and humanly owned. We identify four mechanisms that widen this gap—predictive emotional surveillance, synthetic emotionality, empathy inflation and relational displacement—and show how each reshapes emotional labour and leader–employee relationships. Building on these mechanisms, the paper outlines governance and design strategies that position AI as assistive rather than substitutive, safeguard emotional privacy and deliberately reconnect employees to human support. The argument reframes empathetic leadership as a socio-technical practice in AI-rich workplaces.

DOI: https://doi.org/10.54660/IJMOR.2026.5.4.09-14

Keywords: AI-enabled empathy, empathetic leadership, emotional AI, authenticity in leadership, psychological safety, algorithmic management, workplace surveillance

Introduction

Empathy has become a central expectation of contemporary leadership, particularly in hybrid and digital-first workplaces where emotional cues are harder to read. Empathetic leaders are expected to understand employees' experiences, respond to distress and create climates of psychological safety and inclusion (Brown & Treviño, 2022; Gentry *et al*., 2020) [4, 8]. At the same time, organisations are increasingly adopting AI systems that detect, infer or simulate emotional states, ranging from sentiment analysis and voice stress detection to wellbeing analytics and AI-drafted "empathetic" messages. These technologies are typically framed as tools for augmenting empathetic leadership: they promise earlier detection of burnout, more tailored check-ins and more consistent care at scale. Yet critical scholarship on affective computing, workplace surveillance and emotional AI suggests that such tools may also introduce new risks, including privacy violations, emotional manipulation and erosion of authentic relationships (Kayas *et al*., 2023; Roemmich *et al*., 2023; Babu *et al*., 2025) [13, 24, 1]. This paper examines the dark side of AI-enabled empathy in leadership. Drawing on and extending two appendices that synthesise recent literature on emotional AI and governance responses (Appendix 1 and Appendix 2), it develops the concept of an algorithmic empathy–authenticity gap: a widening disconnects between the quantity, precision and personalisation of empathetic signals and employees' perception that these signals are sincerely felt and humanly owned.

The paper proceeds as follows. Section 2 reviews work on empathetic leadership and emotional AI. Section 3 introduces the algorithmic empathy–authenticity gap. Section 4 presents the Dark-Side Mechanisms of AI-Enabled Empathy. Section 5 discuss governance and mitigation strategies. Section 6 discusses Empathy as a Socio-Technical Construct, and Section 7 concludes the article.

2. Literature Background

2.1. Empathy and leadership authenticity

Empathy in leadership is commonly defined as the ability to understand and share others' emotional states, engage in perspective-taking, and respond in a caring, constructive way (Gentry *et al*., 2007). Empathetic leadership has been associated with higher engagement, trust, psychological safety, and ethical conduct, with follower perceptions of leader integrity and moral character playing a central role in these effects (Brown & Treviño, 2006) [3]. Critically, employees evaluate not only what leaders say but also whether they mean it; authenticity therefore becomes central to the perceived value of empathic behaviour. When empathy is perceived as instrumental, scripted, or performative rather than sincere, it can backfire—elevating cynicism and reducing trust in leadership (Brown & Treviño, 2006) [3].

2.2. Emotional AI and affective computing at work

Affective computing and emotional AI technologies attempt to detect or infer affective states from facial expressions, tone of voice, text patterns, physiological signals, and other behavioural traces. In organisational contexts, such tools are increasingly deployed in customer service, call centres, driver and operator monitoring, employee wellbeing, performance management, and recruitment. Empirical research indicates that employees often experience emotion AI as a form of "deep surveillance" that intrudes into mental and emotional life (Roemmich *et al*., 2023) [24]. The scope of workplace surveillance has expanded from task performance toward behaviour and affect, creating persistent tensions between efficiency and employee dignity (Kayas *et al*., 2023) [13]. Moreover, employees interpret the use of emotion AI through a relational ethics lens, emphasising concerns related to fairness, consent, and potential harm (Corvite *et al*., 2023) [6].

2.3. Human–AI interaction, artificial intimacy and pseudo-empathy

Beyond the workplace, studies of conversational agents and emotional AI companions document new forms of "artificial intimacy," in which users develop perceived emotional relationships with non-sentient systems (Jones, 2025) [11]. Emotional AI may also foster pseudo-intimacy, where patterned responsiveness is interpreted as care rather than as automated interaction (Babu *et al*., 2025) [1]. Experimental evidence further indicates that AI-generated empathetic responses can be rated as more responsive or compassionate than those produced by human experts in certain contexts (Ovsyannikova *et al*., 2025) [22]. Collectively, this work suggests that emotional AI can convincingly simulate empathy cues despite lacking subjective experience and moral accountability (Babu *et al*., 2025; Jones, 2025; Ovsyannikova *et al*., 2025) [1, 11, 22].

3. The Algorithmic Empathy–Authenticity Gap

Beyond the workplace, studies of conversational agents and emotional AI companions document new forms of "artificial intimacy," in which users develop perceived emotional relationships with non-sentient systems (Jones, 2025) [11]. Emotional AI may also foster pseudo-intimacy, where patterned responsiveness is interpreted as care rather than as automated interaction (Babu *et al*., 2025) [1]. Experimental evidence further indicates that AI-generated empathetic responses can be rated as more responsive or compassionate than those produced by human experts in certain contexts (Ovsyannikova *et al*., 2025) [22]. Collectively, this work suggests that emotional AI can convincingly simulate empathy cues despite lacking subjective experience and moral accountability (Babu *et al*., 2025; Jones, 2025; Ovsyannikova *et al*., 2025) [1, 11, 22].4. Dark-Side Mechanisms of AI-Enabled Empathy

4. Dark-Side Mechanisms of AI-Enabled Empathy

4.1. Predictive emotional surveillance

Predictive emotional surveillance refers to the use of emotional AI to continuously infer employees' affective states from digital traces such as voice, facial expression, keystrokes, or chat logs. As summarised in Appendix 1, such tools can integrate emotion scores into HR dashboards and enable always-on mood tracking within collaboration platforms. Empirical work suggests that employees often experience emotion AI as privacy-invasive and as an extension of workplace surveillance into inner emotional life (Roemmich *et al*., 2023; Urquhart *et al*., 2022) [24, 26]. Such perceptions can undermine psychological safety, as employees may mask emotions, disengage cameras, or withdraw from digital channels to reduce the risk of misinterpretation (Roemmich *et al*., 2023; Urquhart *et al*., 2022) [24, 26]. Even when framed as "wellbeing support," these systems can amplify perceived power asymmetries—leaders gain emotional visibility while employees retain limited control over how emotional inferences are generated and used—thereby eroding trust (Roemmich *et al*., 2023; Urquhart *et al*., 2022) [24, 26].

4.2. Synthetic emotionality

Synthetic emotionality refers to situations in which AI composes messages that appear empathic but are not grounded in a leader's own emotional engagement. As summarised in Appendix 1, this can include LLM-assisted drafting of "empathetic" emails, speeches, and performance-review comments, as well as chatbots responding in a leader's name and auto-generated apology messages. Disclosure of AI authorship in emotionally resonant communication has been shown to reduce perceived authenticity and trust (Kirkby, 2023) [14]. Related evidence indicates that messages initially evaluated as high-quality are subsequently downgraded when recipients learn they were AI-generated (Dorigoni & Giardino, 2025). Moreover, some moral and relational acts—such as apologising or expressing gratitude—are arguably "second-personal," meaning their ethical significance depends on who performs them within the relationship (Battisti, 2025) [2]. Consequently, when leaders offload such acts to AI, synthetic emotionality can produce an authenticity violation and may be experienced as emotional manipulation (Battisti, 2025; Dorigoni & Giardino, 2025; Kirkby, 2023) [2, 7, 14].

4.3. Empathy inflation

Empathy inflation occurs when AI-driven wellbeing analytics and nudging systems push leaders toward high-frequency, hyper-personalised displays of care (e.g., "check on X now; sentiment dropped"). As indicated in Appendice 1, wellbeing analytics, micro-targeted outreach suggestions, and systems that reward frequent "supportive touches" can collectively elevate expectations of managerial emotional availability. Evidence from algorithmic management research suggests that sustained micro-prompts and

continuous digital oversight can contribute to technostress, intensification of work demands, and cognitive overload (Mbare *et al*., 2024; Nilsson *et al*., 2025) [16, 20]. When combined with findings that AI-generated empathetic responses may be perceived as more compassionate than those produced by expert humans in certain contexts (Ovsyannikova *et al*., 2025) [22], empathy inflation may raise both employee expectations and leaders' emotional burden, shifting empathy from a meaningful, context-sensitive behaviour to a high-volume performance requirement (Mbare *et al*., 2024; Nilsson *et al*., 2025; Ovsyannikova *et al*., 2025) [16, 20, 22].

4.4. Relational displacement

Relational displacement describes how AI agents become the primary interface for emotional disclosure and support, thereby displacing direct leader–employee relationships. As indicated in Appendix 1, mental-health or HR chatbots, asynchronous AI-mediated feedback, and avatars representing leaders can collectively position AI as the main "listener" in everyday organisational life. Emotional AI companions can foster pseudo-intimacy, encouraging users to prefer frictionless interactions with AI over complex engagements with humans (Babu *et al*., 2025; Jones, 2025) [1, 11]. In organisational settings, this may leave leaders reliant on aggregate sentiment dashboards while employees increasingly confide in systems rather than in people. Over time, such displacement can weaken relational warmth and mutual understanding and shift trust away from interpersonal relationships toward platforms and automated support channels (Babu *et al*., 2025; Jones, 2025) [1, 11].

5. Governance and Mitigation Strategies

Building on the dark-side mechanisms outlined above, Appendix 2 consolidates existing scholarship into a set of governance levers that can help organisations preserve empathy, authenticity, and trust in AI-mediated leadership contexts. It organises governance responses to AI-enabled empathy risks across five levers—capability building, policy and ethics, data governance, system design, and leadership development. For delegated empathy and synthetic emotionality, it prioritises using AI as assistive (not substitutive), strengthening leaders' empathic capability, and treating AI outputs as prompts for real dialogue while keeping leaders accountable for core relational acts (Brown & Treviño, 2006; Butaney & Wortman Vaughan, 2023; Gentry *et al*., 2007) [3, 5]. For predictive emotional surveillance, it stresses proportionate emotional-data policies (opt-in where feasible, data minimisation, separation of wellbeing analytics from performance decisions, and transparent opt-out/contestability) to protect trust and psychological safety (Kayas *et al*., 2023; Milanez *et al*., 2025; Pasquale, 2023; Roemmich *et al*., 2023; Urquhart *et al*., 2022) [13, 18, 23, 24, 26], including worker voice and the ability to contest or withdraw from emotion-AI deployments (Corvite *et al*., 2023) [6]. For synthetic emotionality, it recommends disclosure rules, "human-only" zones (e.g., apologies and sensitive feedback), and requirements that leaders personalise and own AI-assisted drafts to avoid perceived deception (Butaney & Wortman Vaughan, 2023; Dorigoni & Giardino, 2025; Battisti, 2025; Kirkby, 2023) [5, 7, 2, 14]. To prevent empathy inflation, it advises calibrated nudging (lower alert frequency), workload design, and boundary-setting/referral training so empathy is a shared organisational responsibility rather than a 24/7 managerial obligation (Kayas *et al*., 2023; Mbare *et al*., 2024; Nilsson *et al*., 2025; Ovsyannikova *et al*., 2025) [13, 16, 20, 22]. Finally, to counter relational displacement, it calls for "reconnection-by-design" (clear escalation paths to humans) and leader capability to convert AI-mediated touchpoints into genuine conversations (Babu *et al*., 2025; Butaney & Wortman Vaughan, 2023; Milanez *et al*., 2025; Ovsyannikova *et al*., 2025) [1, 5, 18, 22].

6. Discussion: Empathy as a Socio-Technical Construct

This analysis reframes empathetic leadership as a socio-technical construct in AI-rich workplaces. Empathy is not merely a communication skill; it is judged as a signal of authentic intent, integrity, and moral character (Brown & Treviño, 2006; Gentry *et al*., 2007) [3]. At the same time, workplace emotion AI is often experienced as intrusive and asymmetrical, extending managerial visibility into employees' inner emotional life (Kayas *et al*., 2023; Roemmich *et al*., 2023; Urquhart *et al*., 2022) [13, 24, 26]. Integrating these streams shows why emotional AI is not neutral "support": it reshapes how empathy is produced, interpreted, and governed in organisations.

The mechanisms in Appendices 1 and 2 point to three shifts. First, empathy becomes co-produced by leaders and systems—dashboards, nudges, and generative text amplify empathic signals, yet may be attributed to automation rather than genuine concern, widening an algorithmic empathy–authenticity gap. Second, emotional labour is redistributed: AI may reduce some cognitive load but can intensify relational demands by accelerating expectations of responsiveness and "always-on" care (Mbare *et al*., 2024; Nilsson *et al*., 2025) [16, 20]. Third, authenticity becomes contested; when care appears scripted or emotions appear monitored, empathic signals may be read as surveillance or manipulation rather than support (Corvite *et al*., 2023; Roemmich *et al*., 2023) [6, 24]. These concerns are heightened because AI-generated empathic responses can be highly convincing—sometimes judged as more compassionate than expert human messages—raising the stakes for disclosure and ownership (Ovsyannikova *et al*., 2025) [22].

Accordingly, the core issue is legitimacy, not only technical accuracy: whether AI-enabled empathy is perceived as respectful, consensual, and humanly owned. Future research should examine interpretive processes—how employees infer leader intent under AI mediation and how disclosure, climate, and prior trust shape reactions (Kirkby, 2023; Dorigoni & Giardino, 2025) [14, 7]; treat emotional AI as a power-sensitive governance issue linked to data access, contestability, and purpose limitation (Kayas *et al*., 2023; Urquhart *et al*., 2022) [13, 26]; and test boundary conditions across sectors and cultures where surveillance norms, managerial distance, or stigma around emotional disclosure may widen the gap. Practically, organisations should treat emotional AI as infrastructure that must be governed to protect dignity and trust, shifting from tool deployment to socio-technical design—transparent policy, minimal and consensual data practices, and leadership capability-building that preserves the irreducibly human work of empathy.

For leaders, the priority is to avoid delegating emotional labour to AI: use tools to spot patterns or draft language, but own empathic judgement and high-stakes relational acts, treat outputs as prompts for inquiry verified through conversation, personalise communication, and set boundaries to prevent "always-on" expectations, using referrals when appropriate.

For organisations, emotional AI should be governed as high-risk infrastructure through purpose limitation, data minimisation, transparency, separation of wellbeing analytics from performance management, disclosure norms, "human-only" zones, capability-building, and monitoring of impacts on trust and psychological safety, with contestability and opt-out where feasible.

7. Conclusion

AI-enabled empathy tools promise to enhance leaders' awareness of employee wellbeing and support more personalised communication in digital and hybrid workplaces. However, this article demonstrates that emotional AI also introduces significant socio-technical risks that can undermine authenticity, trust, and psychological safety. By conceptualising the algorithmic empathy–authenticity gap, the study shows how empathy may become more frequent and data-driven while being perceived as less sincere and less humanly owned. The analysis identifies four dark-side mechanisms—predictive emotional surveillance, synthetic emotionality, empathy inflation, and relational displacement—through which AI-mediated empathy can be reinterpreted as monitoring, manipulation, or performance. These mechanisms highlight that the core challenge is not technical accuracy but relational legitimacy. Addressing this challenge requires governance beyond technical design, treating emotional AI as high-risk infrastructure and anchoring its use in leadership capability, ethical policy, data governance, and human-centred system design. Overall, the findings underscore a clear implication: empathetic leadership must remain human-led, with AI used only as a bounded and transparent support. When governed responsibly, emotional AI can enhance awareness and responsiveness; when uncritically adopted, it risks automating the appearance of care while eroding the human relationships that give empathy its meaning.

References

1. Babu J, Joseph D, Kumar RM, Alexander E, Sasi R, Joseph J. Emotional AI and the rise of pseudo-intimacy: Are we trading authenticity for algorithmic affection? Front Psychol. 2025;16:1679324. doi:10.3389/fpsyg.2025.1679324.
2. Battisti D. Second-person authenticity and the mediating role of AI: A moral challenge for human-to-human relationships? Philos Technol. 2025;38(1):1-19. doi:10.1007/s13347-025-00857-w.
3. Brown ME, Treviño LK. Ethical leadership: A review and future directions. Leadersh Q. 2006;17(6):595-616. doi:10.1016/j.leaqua.2006.10.004.
4. Brown ME, Treviño LK. Ethical leadership: A review and future directions. J Organ Behav. 2022;43(1):1-19. doi:10.1002/job.2596.
5. Butaney G, Wortman Vaughan J. Responsible AI for people management. AI Soc. 2023;38:789-803. doi:10.1007/s00146-022-01579-5.
6. Corvite S, Roemmich K, Rosenberg TI, Andalibi N. Data subjects' perspectives on emotion artificial intelligence use in the workplace: A relational ethics lens. Proc ACM Hum Comput Interact. 2023;7(CSCW1):Article 124. doi:10.1145/3579600.
7. Dorigoni A. The illusion of empathy: Evaluating AI-generated outputs in moments that matter. Front Psychol. 2025;16:1568911. doi:10.3389/fpsyg.2025.1568911.
8. Gentry WA, Deal JJ, Stawiski SA, Ruderman M. Empathy in the workplace: A tool for effective leadership. Greensboro (NC): Center for Creative Leadership; 2020.
9. Gualeni S. Second-person authenticity and the mediating role of AI: A moral challenge for human-to-human relationships? Philos Technol. 2025. Advance online publication. doi:10.1007/s13347-025-00857-w.
10. Nilsson KH, Matilla-Santander N, Lee MK, Brulin E, Bodin T, Håkansta C. Algorithmic management and occupational health: A comparative case study of organizational practices in logistics. Saf Sci. 2025;187:106863. doi:10.1016/j.ssci.2025.106863.
11. Jones M. Artificial intimacy: Exploring normativity and personalization through fine-tuning LLM chatbots. In: Proceedings of the ACM Conference. New York: Association for Computing Machinery; 2025. doi:10.1145/3706598.3713728.
12. Kayas OG. Workplace surveillance: A systematic review, integrative framework, and research agenda. J Bus Res. 2023;168:114212. doi:10.1016/j.jbusres.2023.114212.
13. Kayas OG, Efthymiadis A, Nguyen MH, Warkentin M. Workplace surveillance: A systematic review, integrative framework, and research agenda. J Bus Res. 2023;168:114212. doi:10.1016/j.jbusres.2023.114212.
14. Kirkby A. To disclose or not disclose is no longer the question: The effect of AI-disclosed brand voice on brand authenticity and trust. J Prod Brand Manag. 2023;32(7):1108-1121. doi:10.1108/JPBM-02-2022-3864.
15. Mbare B. Algorithmic management, wellbeing and platform work: A psychosocial risk perspective. Econ Labour Relat Rev. 2024. Advance online publication. doi:10.1177/10353046241242344.
16. Mbare B, Perkiö M, Koivusalo M. Algorithmic management, wellbeing and platform work: Understanding the psychosocial risks and experiences of food couriers in Finland. Labour Ind. 2024;34(1):1-26. doi:10.1080/10301763.2024.2423442.
17. Menges L, Weber-Guskar E. Digital emotion detection, privacy, and the law. Philos Technol. 2025. Advance online publication. doi:10.1007/s13347-025-00895-4.
18. Milanez A, Lemmens A, Ruggiu C. Algorithmic management in the workplace: New evidence from an OECD employer survey. OECD Artificial Intelligence Papers. No. 31. Paris: OECD Publishing; 2025. doi:10.1787/287c13c4-en.
19. Nilsson KH, Andersson L, Johansson B. Algorithmic management and occupational health: A comparative case study of organisational practices in logistics. Saf Sci. 2025;181:106424. doi:10.1016/j.ssci.2024.106424.
20. Nilsson KH, Matilla-Santander N, Lee MK. Algorithmic management and occupational health: A comparative case study of organizational practices in logistics. Saf Sci. 2025;187:106863. doi:10.1016/j.ssci.2025.106863.
21. OECD. Algorithmic management in the workplace. Paris: OECD Publishing; 2025. doi:10.1787/287c13c4-en.
22. Ovsyannikova D, Li JZ, Perry A. Third-party evaluators perceive AI as more compassionate than expert humans. Commun Psychol. 2025;3:182. doi:10.1038/s44271-024-00182-6.

23. Pasquale F. Affective computing at work: Rationales for regulating emotion attribution and manipulation. Brussels: European Trade Union Institute; 2024. p. 1-22.
24. Roemmich K, Schaub F, Andalibi N. Emotion AI at work: Implications for workplace surveillance, emotional labor, and emotional privacy. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). New York: Association for Computing Machinery; 2023. doi:10.1145/3544548.3580950.
25. Srinivasan R, San Miguel González B. The role of empathy for artificial intelligence accountability. J Responsible Technol. 2022;9:100021. doi:10.1016/j.jrt.2021.100021.
26. Urquhart L, Laffer A, Miranda D. Working with affective computing: Exploring public perceptions of AI-enabled workplace surveillance [preprint]. arXiv. 2022. Available from: arXiv:2205.0826.

How to Cite This Article
Tiruwa A, Dikshit S. The algorithmic empathy–authenticity gap: a perspective on the dark side of AI-enabled empathy in leadership. Int J Manag Organ Res. 2026;5(4):9-14. doi:10.54660/IJMOR.2026.5.4.09-14.

Appendix 1

Dark-side risk mechanisms of AI-enabled empathy in leadership and indicative supporting literature			
Risk mechanism	**Synthesized definition**	**Typical AI practices / technologies**	**Supporting literature**
Predictive Emotional Surveillance: Privacy and safety threats	Emotional AI continuously infers employees' affective states from digital traces. The boundary between care and control blurs, and "empathetic monitoring" is experienced as surveillance. Employees self-censor and emotionally withdraw.	• Affective computing tools, reading voice, facial micro-expressions, keystrokes, or chat logs. • Emotion scores integrated into HR dashboards. • Always-on mood tracking in collaboration platforms.	Urquhart *et al*., 2022; Roemmich, 2023; Pasquale, 2024; Menges, 2025; OECD, 2025
Synthetic Emotionality: AI-generated empathy without human intent	AI composes messages that sound empathetic but are not grounded in a leader's own emotional engagement. When employees discover this, they experience a strong authenticity violation and may interpret the practice as emotional manipulation.	• LLM-based assistants drafting "empathetic" emails, speeches, or review comments. • Chatbots responding in the leader's name. • Auto-generated apology or "I care about you" messages.	Kirkby, 2023; Dorigoni, 2025; Gualeni, 2025
Empathy Inflation: Over-personalisation and emotional dependency	Continuous AI nudges push leaders toward high-frequency, hyper-personalised empathetic gestures ("check on X now"), inflating expectations of emotional availability. Employees come to expect constant care; leaders feel pressured and emotionally over-extended.	• Wellbeing analytics pinging managers with real-time prompts • Micro-targeted "reach out" suggestions based on sentiment spikes • Systems rewarding high volumes of "supportive touches"	Kayas *et al*., 2023; Mbare, 2024; Nilsson *et al*., 2025; Ovsyannikova *et al*., 2025; OECD, 2025
Relational Displacement: AI intermediates the human connection	AI becomes a primary interface for emotional support, displacing direct human–human interaction. Employees may feel more comfortable with AI tools than with leaders, weakening core relational bonds and shared vulnerability.	• Mental-health or HR chatbots positioned as the primary "listener". • Asynchronous AI-mediated feedback instead of direct conversations. • Avatars/agents representing leaders in digital environments.	Kayas *et al*., 2023; Babu, 2025; Ovsyannikova *et al*., 2025; OECD, 2025

Appendix 2

Governance and mitigation strategies for algorithmic empathy risks, with indicative supporting literature			
Risk Mechanism	**Governance focus**	**Key mitigation strategies**	**Indicators of effective mitigation**
Delegated Empathy	• Capability building • Role clarity	Position AI as assistive, not a substitute for human listening and care; treat AI insights as cues for conversation and embed human empathy behaviours in competency models (Brown & Treviño, 2022; Gentry *et al*., 2020; Butaney & Wortman Vaughan, 2023).	• Employee surveys show stable or improved ratings for "my manager genuinely listens to me" (Gentry *et al*., 2020). • Use of AI tools coexists with strong perceptions of leader availability and care (Brown & Treviño, 2022; Butaney & Wortman Vaughan, 2023).
Predictive Emotional Surveillance	• Policy and ethics • Data governance	Adopt clear, proportionate policies for emotional data; minimal, non-biometric analytics; separate wellbeing analytics from performance management and communicate risks and opt-out options (Kayas *et al*., 2023; Urquhart *et al*., 2022; Roemmich, 2023; Pasquale, 2024; OECD, 2025).	• Employees can accurately state the organisation's rules on emotional data and feel they have meaningful choice (Urquhart *et al*., 2022). • No documented use of emotional analytics in ratings, discipline, or dismissal cases (OECD, 2025). • Psychological safety and trust indices remain stable or improve after introducing tools (Kayas *et al*., 2023; Roemmich, 2023).
Synthetic Emotionality	• Design transparency • Communication norms	Disclose substantial AI involvement in emotionally significant messages, reserve "human-only" zones for apologies and sensitive feedback, and require leaders to personalise and own AI-assisted drafts (Kirkby, 2023; Dorigoni, 2025; Gualeni, 2025; Butaney & Wortman Vaughan, 2023).	• Employees describe leader messages as "genuine" and "from the person, not just the system" in qualitative feedback (Kirkby, 2023). • Few or no complaints of feeling deceived when AI involvement is discovered (Dorigoni, 2025). • Perceptions of leader authenticity remain high even as AI tools are adopted (Gualeni, 2025; Butaney & Wortman Vaughan, 2023).
Empathy Inflation	• Workflow and load design • Wellbeing policy	Calibrate nudges so alerts are occasional and triage-oriented, combine AI-prompted outreach with peer and professional supports, and train leaders in boundary-setting and referral practices (Nilsson *et al*., 2025; Mbare, 2024; Kayas *et al*., 2023; Ovsyannikova *et al*., 2025).	• Leaders report manageable cognitive/emotional demands and low technostress related to AI tools (Nilsson *et al*., 2025; Mbare, 2024). • Employees see empathy as a shared responsibility (self, peers, leaders, HR), not something the manager must perform 24/7 (Kayas *et al*., 2023). • Expectations of "always-on empathy" are moderated even as empathic AI remains available (Ovsyannikova *et al*., 2025).
Relational Displacement	• Experience design • Leadership development	Design AI to complement, not replace, human dialogue; build "reconnection to humans" into system flows; and develop leaders' skills to invite and hold deeper conversations after AI-mediated contact (Babu, 2025; OECD, 2025; Butaney & Wortman Vaughan, 2023; Ovsyannikova *et al*., 2025).	• Employees report that AI tools "make it easier" rather than harder to talk with their manager (Babu, 2025; Ovsyannikova *et al*., 2025). • Perceived closeness to leaders and team members remains stable or improves after AI deployment (OECD, 2025). • Data shows healthy rates of escalation from AI to human conversations instead of emotional isolation within the system (Butaney & Wortman Vaughan, 2023).