



Digital Article / Collaboration and Teams

How to Foster Psychological Safety When AI Erodes Trust on Your Team

Proactively addressing team dynamics can prevent costly errors and cycles of dysfunction. *by Jayshree Seth and Amy C. Edmondson*

Published on HBR.org / February 4, 2026 / Reprint [H0924Z](#)



Richard Drury/Getty Images

There may be something unsettling happening on your team. Despite expected productivity gains from integrating AI tools, overall team performance appears to be declining. People are starting to second-guess themselves, and trust is eroding in ways that are hard to pinpoint.

Consider this hypothetical scenario: Your marketing team uses a gen AI tool to analyze campaign performance. The AI confidently recommends

reallocating budget based on what appears to be solid data analysis. But it turns out the recommendation is based on a fundamental misunderstanding of your customer segments. The campaign fails, budgets are wasted, and suddenly everyone on the team is questioning not just the AI's future recommendations, but their own ability to evaluate them.

As organizations rapidly integrate AI into their workflows, leaders are discovering that the challenges it brings aren't just technical—they're fundamentally about team dynamics as well. The same AI tools that promise to enhance productivity can create predictable patterns of team dysfunction that mirror classic organizational behavior problems. Consider the rise in "workslop": AI-generated output that fails to move a project forward and dumps extra cognitive and emotional labor on colleagues who must fix or redo it. The result is not only harm to productivity but damaged trust in coworkers.

But here's what's different: Most leaders don't recognize these as team integration challenges. They may treat them as technology problems to be solved with better tools or training, rather than team effectiveness issues that require many of the principles we know work for human collaboration.

In our work—Jayshree developing gen AI use cases for R&D at 3M and Amy reviewing AI implications for the workforce and society—we've found that understanding human-AI team dynamics is crucial not only from an organizational theory and practice standpoint, but for preventing costly errors and the erosion of team trust.

Where Trust Breaks Down

Trust is known to be critical to team performance, and AI can change that dynamic. When AI provides confident but incorrect

information, it creates “trust ambiguity.” Ambiguity exists when people believe trust is supposed to be warranted while lacking actual trust; further, that lack of trust often feels undiscussable. In a state of trust ambiguity, team members lose confidence not just in the AI’s output, but in their own judgment. Research shows that sustained AI use undermines professionals’ confidence in their ability to challenge AI recommendations, even when they possess the expertise to do so. All this in turn directly threatens the psychological safety that’s foundational to team learning and performance, because ambiguity inhibits willingness to speak up.

Consider another hypothetical scenario where AI confidently recommends pursuing a “breakthrough” sustainable technology based on patent analysis but misinterprets expired patents as active innovations, leading to investment discussions in an obsolete technology.

This isn’t just about AI’s accuracy; it’s about how AI’s inaccuracies ripple through teams in ways that human mistakes don’t. When a human colleague makes an error, teams often engage in collective sense-making and update their mental models and understanding of who knows what.

Humans can ask questions (“What data were you looking at?” “What was your reasoning?”), understand the context (“You were rushing to meet the deadline”), and collaborate on prevention (“Let’s build in a review step”). This process of mutual learning and accountability exploration can actually strengthen team learning and bonds.

Gen AI errors can short-circuit this fundamental team learning process. When AI confidently makes a recommendation, teams can’t engage in collaborative sense-making due to the “black box” nature of AI.

This lack of transparency and explainability can make team members struggle to calibrate trust in AI and feel confidence in its output.

The team may move forward, trusting AI and shorting its normal checks and discussions, leading to missteps. They can't challenge AI on its assumptions, check its methodology, or understand the chain of reasoning used. This creates attribution uncertainty—teams know something went wrong due to the AI output but have no clear pathway to understand what to attribute it to or how to reduce uncertainty or prevent its recurrence.

This uncertainty can then compound further, leading to trust ambiguity. Teams can begin questioning other AI outputs (“If it was wrong about this, what else is it wrong about?”) and eventually their own judgment (“Are we being too trusting of AI? Too skeptical? How do we calibrate?”). Unlike human errors, which, when surfaced, get chewed, digested, ruminated, and metabolized through team discussion, AI errors can create expanding circles of doubt that have no clear path to resolution.

Where Coordination Breaks Down

Recent studies have shown that having an AI team member can negatively impact important team processes by disrupting team dynamics. For example, in one study, teams with AI members showed increased coordination problems and lower performance due to reduced effort among the human teammates.

This reveals a common challenge: The presence of AI on a team might impair important group processes, such as communication and coordination, in addition to effort and accountability, resulting in less-effective interactions among team members.

This aligns with what's known for organizational behavior: Effective teaming requires the fluid, adaptive collaboration that emerges when team members recognize the need to anticipate each other's needs and behaviors. AI disrupts this process in subtle but significant ways, impacting both trust development and decision-making. Unlike human teammates, AI doesn't pick up on contextual cues, doesn't adjust its communication style based on team dynamics, and can't connect and engage in the informal relationship-building that makes teams effective. AI also doesn't suffer consequences for being wrong. There can be an accumulated cost of having to work around a "team member" that operates according to fundamentally different rules.

Moreover, given its perception as a technology with access to seemingly unlimited information, human team members may become overconfident in AI's capabilities, cognitively offloading as a result. This can lead to team members relaxing their own sense of accountability and critical engagement, creating a "human-AI oversight paradox."

Applying Psychological Safety Principles to AI Integration

The solution isn't to abandon AI or accept these shortcomings as inevitable. Instead, leaders need to apply proven organizational behavior principles to AI integration and related change management. This requires treating AI integration as fundamentally a *team learning challenge*, not just a technology implementation one. Here's how:

Reframe AI integration as a learning process, not an execution process.

Just as psychological safety is built by framing work as a learning task rather than an execution task, successful AI integration requires positioning AI deployment as ongoing experimentation with hands-on learning. Leaders should communicate that they expect the team will discover AI limitations and that, as in any new frontier, this is valuable intelligence, not failure. Early AI "mistakes" should be treated

as learning opportunities that help the team calibrate expectations and develop better collaboration protocols.

Instead of asking, “Is the AI working well?” leaders should ask team members, “How is working with AI affecting how you collaborate with each other?” and “What are you learning about when to trust AI versus when to rely on human judgment?” Reward the right behaviors by celebrating team members who catch AI errors rather than those who blindly accept AI outputs. Make questioning AI a sign of good judgment, not resistance to innovation.

In developing use cases for gen AI in R&D at 3M, rather than mandating adoption after rolling out a tool, the team invited volunteers to participate in early pilots. The goal was to clearly signal that they were in a learning mode, not deployment mode. Volunteers and the wider organization knew that they could help them in understanding what works, not just evaluating a finished product they created. The team specifically highlighted improvements based on user feedback to demonstrate a learning mindset in action. Moreover, instead of simply talking in terms of time saved or tasks automated, they also relied on feedback and testimonials to inspire confidence among others.

Model fallibility and curiosity.

One of the ways leaders build psychological safety is acknowledging their own fallibility. With respect to AI, this means openly admitting when you don’t understand AI outputs or when they seem questionable. Leaders should encourage responsible use of AI and share their own AI mistakes and what they learned, demonstrating that questioning AI is not only acceptable but expected. Relatedly, they should model curiosity about how and why AI works and will work going forward, so the downsides can be better understood and the upsides better realized. Consider instituting learning rituals like regular “AI After-

Action Reviews” where teams discuss what worked, what didn’t, and why.

At 3M, the team routinely posted in internal channels about new learnings, unexpected insights, and AI limitations encountered. The goal of this reframing was to change how the technical community perceived the effort, allowing them to clearly see that the team charged with developing tools was engaged in actively learning alongside them. They did much of the sharing against the backdrop of demystifying the AI black box, and as more technical employees understood that AI uses pattern matching, not “thinking,” they became more curious about its limitations as opposed to blindly trusting or outright rejecting it. The team also conducted many sessions that modeled AI curiosity rather than justification of the use of AI.

Create intelligent failure protocols.

Drawing on principles from Amy’s book, *Right Kind of Wrong*, teams need to distinguish between different types of AI failures. “Intelligent AI failures”—testing AI in new domains or pushing its boundaries in controlled ways where the risks are low—should be celebrated as learning opportunities. “Basic AI failures” like not verifying AI outputs in domains where limitations are already known should be prevented through better processes.

3M therefore initially tested tools with one to three technical experts tasked with reporting every issue they could find. When the experts identified problems, they shared them widely and showed exactly what changes were instituted—creating a visible failure-to-improvement loop. The 3M team also expanded thoughtfully: first with a small, focused expert group, then with a larger pilot with volunteers and wider local deployment, followed by plans for regional and global rollout to build multiple checkpoints to catch and learn from failures.

Emphasize human connection.

Perhaps most critically, successful AI integration requires intentionally preserving human connection. As AI handles more routine tasks, the remaining human work will become more complex and interdependent. This makes psychological safety more important, not less. Leaders must create explicit space for human-only discussions about team dynamics and AI integration challenges. They need to recognize and discuss team members' concerns about AI replacing human value, emphasizing how AI enhances rather than diminishes human expertise.

Moreover, anthropomorphizing AI to make it more trustworthy can have unintended consequences, creating unrealistic expectations about its capabilities. Current AI systems aren't really intelligent but instead provide "artificial diligence" that can assist and augment human capabilities; yet most people, when surveyed, think of AI as a tool for problem-solving.

Leaders should also create override protocols to ensure that team members can easily override AI recommendations without having to justify their decisions extensively. Sometimes human intuition catches what AI analysis can miss.

Enabling the Team to Thrive

Trust is integral to human-AI teaming. Teams must operate in an environment where everyone feels able to question and challenge AI's performance. People must feel comfortable admitting confusion about AI recommendations, reporting AI errors, suggesting improvements to human-AI workflows, and expressing concerns about AI replacing human judgment. Leaders need to acknowledge and address the distinct ways in which AI can disrupt normal team functioning, from the opacity that prevents questioning assumptions to the need

for ongoing learning about AI's rapidly evolving capabilities and limitations.

The future of work isn't about choosing between human intelligence and artificial intelligence—it's about building teams that allow both to contribute to their fullest potential. Importantly, AI integration success must be measured not just by AI performance metrics, but by overall team effectiveness, learning velocity, and the ability to leverage both human and artificial intelligence optimally. But this requires treating AI integration as an effective team development challenge, not just as efficient technology deployment. Just as we've learned that effective teaming requires skills that may not come naturally to all, effective human-AI collaboration requires new competencies that organizations must deliberately develop.

The good news is that to minimize trust ambiguity, many of the principles that work for human teams—psychological safety, clear communication, and continuous learning from intelligent failure—can also work for human-AI teams. They just require more intentional and intelligent application.

This article was originally published online on February 4, 2026.



Jayshree Seth is the chief science advocate at 3M, where she leads use cases for gen AI. She is also the author of *The Heart of Science* book trilogy, a TEDx speaker, and was named to the 2025 Thinkers50 Radar.



Amy C. Edmondson is the Novartis Professor of Leadership and Management at the Harvard Business School and the author of *Right Kind of Wrong: The Science of Failing Well*. She is an expert on team learning and psychological safety.