

DISSERTATION PROPOSAL

Allen S. Brown

“Psychological Safety in the Age of Generative AI”

Thursday, December 4, 2025

10:00am

Tepper 4242

Chapter 1: Creating Psychological Safety: An Integrative Review of Antecedents in the Age of Generative AI

Collaborative work inherently involves interpersonally risky exchanges. Psychological safety, the shared belief that it is safe to take interpersonal risks at work, is one of the most influential constructs in organizational behavior used to conceptualize these exchanges. Extensive evidence links psychological safety to learning, voice, innovation, and performance. Yet despite its maturity, recent syntheses highlight a core limitation: scholars still lack a coherent, causal account of how psychological safety is created. In response, this integrative review systematically consolidates and reinterprets research that treats psychological safety as an outcome, drawing on work across organizational behavior, social psychology, communication, and human–technology interaction. Advancing this literature, I develop a dual-path framework in which the emergence of psychological safety is shaped by how individuals appraise threat and learning opportunity within a given interaction space. This framework explains how individual, relational, and technological antecedents jointly influence perceptions of interpersonal risk and possibility, thereby driving the dynamic and emergent formation of psychologically safe climates. I then apply this framework to reassess classic antecedents in the context of generative AI, examining how AI-mediated communication, feedback, and supervision reshape threat and learning appraisals in ways that may either foster or undermine psychological safety. I conclude by articulating a research agenda that integrates human, relational, and technological perspectives to advance a deeper causal understanding of how psychological safety emerges—and evolves—in increasingly AI-mediated workplaces.

Chapter 2: Safe to Talk: Psychological Safety and Generative AI at Work

Generative AI (GenAI) introduces new dynamics into collaborative work by blurring the line between technological tool and social partner. This shift raises theoretical questions about the psychological conditions that enable learning and collaboration in digitally mediated contexts. In this chapter I examine how GenAI’s association with evaluation influences perceptions of autonomy and, in turn, psychological safety. Across two preregistered experiments, participants communicated with either a human or GenAI partner while completing tasks framed as evaluative or non-evaluative. Results show that in non-evaluative contexts, psychological safety does not differ between GenAI and human partners. However, when communication with a work partner contributes to evaluation—a context I describe as indirectly evaluative—psychological safety declines only for GenAI partners. This effect is mediated by reduced perceptions of autonomy, and an intervention that directly affirms autonomy overcomes these differences. These findings extend theory of psychological safety to algorithmically mediated collaboration and identify autonomy as a causal mechanism shaping interpersonal climate in such contexts. Together, this work identifies the value of affirming perceptions of individual autonomy to help preserve the psychological safety necessary for open communication, creativity, and sustained learning when implementing GenAI systems at work.

Chapter 3: Using Chat-based AI to Actively Facilitate Psychologically Safe Climates

Chapter 3 extends the work in chapter two by examining how AI tools might move beyond avoiding harm and towards actively facilitating psychologically safe climates in collaborative work. I advance the dual-pathway theory in which psychological safety emerges from two interconnected mechanisms: reducing interpersonal threat and enhancing collaborative learning opportunities. Drawing on job demands-resources

and stress appraisal theories, I argue that psychological safety is strongest when teams experience productive friction, marked by low threat and high learning. This framework explains why efforts to build psychological safety often fail; interventions that improve one pathway frequently undermine the other. AI tools, however, may offer distinctive capabilities for navigating this tension. Chat-based AI can reduce threat by providing non-judgmental, socially risk-free avenues for inquiry while simultaneously enhancing learning through rapid information access, diverse perspectives, and iterative feedback. However, implementations must also avoid the introduction of new threats associated with monitoring concerns that reduce perceptions of autonomy and hinder learning through shallow responses. I outline three complementary studies to test these dynamics. Study 1 experimentally validates the dual-pathway framework by manipulating threat and learning independently. Study 2 compares psychological safety in AI-assisted versus human-facilitated teams. Study 3 tests AI-based interventions in an organizational field setting. Together, these studies clarify how AI transforms the interpersonal conditions that enable collaborative learning and innovation.

Proposed Committee: Anita Williams Woolley (Chair), Denise M. Rousseau, Lindsay Larson, Jean-François Harvey (Outside Reader)