ISSN: 3068-9503

Research Article | AI and Intelligent Systems: Engineering, Medicine Society

A Multi-Agent System Approach to Psychological Safety in Diverse Teams: An Organizational Behavior Perspective

Wenzhe Song*

School of Business Stevens Institute of Technology, 1 Castle Point Terrace, Hoboken, NJ 07030, USA

***Corresponding Author**
Wenzhe Song, School of Business Stevens Institute of Technology, 1 Castle Point Terrace, Hoboken, NJ 07030, USA.

Submitted: 2025, Oct 10 ; **Accepted:** 2025, Nov 28; **Published:** 2025, Dec 10

Citation: Song, W. (2025). A Multi-Agent System Approach to Psychological Safety in Diverse Teams: An Organizational Behavior Perspective. *AI Intell Sys Eng Med Society, 1*(2), 01-07.

Abstract

In modern organizational behavior and design research, implying machine learning techniques have become the major computational approach, machine learning can excel tasks with labeled data. However, they cannot explain why these patterns emerge, are not equipped to test unknown interventions, and are incapable of exploring causal mechanisms underlying organizational phenomena. In this study, we introduce a multi-agent system approach to simulate 2160 teams and we identify psychological safety is the main factor of diversity's impact. The team's performance reached a peak when diversity ranges from 0.68–0.72, it delivered a massive 34% gain under the condition that psychological safety surpasses 0.53. Reversing its effect from negative (r = 0.34) to positive (r = 0.52). The performance improved 38% despite the conflict having decreased 64%. Especially the early-stage behavioral shaping is 3.8 times more effective than delayed interventions. Behavioral patterns: Empathic mirroring amplifies psychological safety (+0.23), while exclusion mitigates it (0.31). Multi-agent system discloses the fundamental structure psychological safety lessens diversity's organization costs via re- fined communication channels. The model's validation (r = 0.71) confirms the robust predictive power, and optimized teams operate 47% better these insights have not been discovered by machine learning frameworks.

Keywords: Multi-Agent Systems, Organizational Behavior and Design, Psychological Safety, Team Diversity, Agent-Based Modeling

1. Introduction

1.1. Theoretical Background

Recent research in organizational behavior signals a conflict: while diverse teams display stronger innovative potential, they regularly yield lower outcomes than homogeneous ones caused by communicative discrepancies, unclear hierarchical positioning, and dismantled coordination. term this the "diversity paradox," determining psychological safety as the main influencing factor [1]. In their study of 62 pharmaceutical R&D teams, only individuals with high psychological safety were able to fully realize the benefits of diversity, through mechanisms such as framing, inquiry, and boundary-bridging. Drawing from this insight, present a large-scale study of over 4,000 teams, offering a practical framework for modern cross-domain alliance [2]. Their work reinforces regulated enactments (e.g., retrospectives), instantaneous detection of interactional shifts within the team, and adaptive realignment, especially important in remote and hybrid settings. Teams applying these strategies saw client impact metrics rise by up to 23%, with optimal collaboration at 50% in-person time.

Yet psychological safety is degraded by concrete behavioral patterns. Recognize "detractor behaviors" such as ignoring, gaslighting, and code-switching that re- duce inclusion and unevenly harm neglected voices [3]. Their framework affirms the value of exchanging these patterns with planned "amplifier behaviors" to build a culture of habitual inclusion. A related but understudied

factor is "status intelligence" the ability to sense informal rankings and social influence. Show that teams with high-status intelligence members experience fewer conflicts and achieve massively better outcomes [4]. This capacity is especially vital in inclusive teams, where social indicators and governance systems are often vague or misjudged. Finally, warn that internal team competition can improve motivation but decrease innovation by limiting information deficiency [5]. Their results on "bridge teams" which share across boundaries suggest that collaborative network structures improve innovation by nearly 50%, illustrating the value of open, trust-based coordination.

1.2. Research Gap and Contribution

Although previous studies offer valuable insights into distinct drivers of team performance, they usually evaluate these mechanisms in separation. Current research does not provide a coordinated model that represents how psychological safety, status intelligence, inclusive behaviors, measure- ment routines, and network structures interact adaptively over time to impact team outcomes. Moreover, existing data-driven methods, including statistical modeling and machine learning, are limited in their ability to examine temporal, causal, and counterfactual dynamics. They lack the capacity to:

- Model feedback loops between individual behaviors and emergent team climate
- Identify nonlinear thresholds, such as the point at which psychological safety reverses the effect of diversity
- Test intervention timing and sequencing under controlled, repeatable conditions
- Examine how psychological safety diffuses across varying team network structures

To address these limitations, we introduce a Multi-Agent System (MAS) model following that combines the five dimensions into an integrated computational frame- work [6]. The model simulates agents with varying levels of status intelligence engaging with inclusive or selective behaviors within time-varying network topologies. This allows us to uniformly ex- amine how psychological safety emerges, declines, and spreads across teams—and how different intervention strategies affect the relationship between diversity and performance.

This MAS-based approach contributes to literature in three ways:
1. It enables causal modeling of complex team dynamics through mechanism-based modeling.
2. It provides a platform for simulation-based testing, allowing us to test policy measures and behavioral sequences on a scale.
3. It generates testable, data-driven predictions that can frame future research and organizational practice.

2. Methodology

2.1. Overview

We develop a multi-agent system to model team dynamics, focusing on how psychological safety mediates the relationship between diversity and performance. Our framework integrates individual agent characteristics, interaction dynamics, behavioral patterns, and intervention strategies within a unified computational model.

2.2.Why MAS, Not Machine Learning

In organizational behavior, many scholars still rely on machine learning (ML) to study team per- formance. However, ML only detects patterns; it cannot explain why outcomes occur, nor simulate what-if scenarios [7]. For dynamic and emergent phenomena like psychological safety (PS), multi-agent systems (MAS) are more suitable [8,9]. ML treats individuals as data points; MAS treats individuals as interacting agents, each with internal states, decisions, and learning capabilities. This allows us to simulate how PS evolves from many micro-level interactions, and how interventions affect system outcomes before real-world deployment [10,11].

2.3. Core Definitions

2.3.1. Agent State

Each agent i at time t is characterized by:

$$\mathbf{s}_i(t) = (\mathbf{d}_i, \boldsymbol{\theta}_i, \mathbf{x}_i(t)) \tag{1}$$

where:

- $\mathbf{d}_i \in \{0, 1\}^k$: demographic attributes (gender, ethnicity, expertise, age, tenure)
- $\boldsymbol{\theta}_i \in [0, 1]^4$: psychological traits [status intelligence, openness, trust, resilience]
- $\mathbf{x}_i(t) \in [0, 1]^3$: dynamic state [psychological safety ψ_i, performance ρ_i, stress σ_i]

2.3.2.Team Metrics

Team diversity is measured using Simpson's Index:

$$d = 1 - \frac{1}{K}\sum_{k=1}^{K}\sum_{j}\left(\frac{n_{kj}}{N}\right)^2 \tag{2}$$

where n_{kj} is the count of agents with attribute j in dimension k. Average psychological safety is:

$$\bar{\psi} = \frac{1}{N}\sum_{i=1}^{N}\psi_i \tag{3}$$

2.4. Interaction Dynamics

2.4.1. Contact and Quality

The probability of interaction between agents i and j is:

$$p_{ij} = \frac{w_{ij}}{1 + d_{ij}} \cdot \sigma(\boldsymbol{\beta}^T \mathbf{f}_{ij}) \tag{4}$$

where:

$w_{ij} \in [0, 1]$: connection strength in the network

- d_{ij}: network distance between agents
- $\mathbf{f}_{ij} = [1, \kappa_{ij}, \bar{\psi}, -\Delta s_{ij}, \bar{\tau}]^T$: feature vector

- $\kappa_{ij} = e^{-\|d_i - d_j\|/\lambda}$: similarity measure

2.4.2. State Evolution

Psychological safety evolves according to:

$$\psi_i(t+1) = (1-\gamma)\psi_i(t) + \alpha \sum_j w_{ij} q_{ij} + \beta \ell(t) + \epsilon_i \quad (5)$$

where:
- $q_{ij} \in \{-1, 0, 1\}$: interaction quality
- $\ell(t) \in [0, 1]$: leadership influence
- γ: decay rate
- α, β: influence parameters
- $\epsilon_i \sim N(0, \sigma^2)$: stochastic variation

2.5. Performance Model

2.5.1. Team Performance

Overall team performance is:

$$P = \mathbf{w}^T [\bar{\psi}, I, \bar{\rho}, \nu]^T \quad (6)$$

where:
- $I = \rho(G) \cdot \langle BC \cdot \theta_1 \rangle$: information flow (network density × average centrality-weighted status intelligence)
- $\nu = d \cdot \bar{\psi} \cdot \text{ideas}/t$: innovation metric
- $\bar{\rho}$: average individual performance

2.5.2. Diversity-Performance Relationship

The relationship between diversity and performance depends on psychological safety:

$$P(d, \bar{\psi}) = \alpha(d) + \beta(d, \bar{\psi}) \cdot \mathbb{1}[\bar{\psi} \geq \psi^*] + \epsilon \quad (7)$$

where:

$$\alpha(d) = \alpha_0 + \alpha_1 d + \alpha_2 d^2 \quad (8)$$

$$\beta(d, \bar{\psi}) = \beta_1 d \bar{\psi} \quad (9)$$

The critical psychological safety threshold is:

$$\psi^* = \phi_0 + \phi_1 d + \phi_2 \log N + \phi_3 d \log N \quad (10)$$

2.6. Behavioral Patterns

We model behavioral patterns using a unified framework:

$$P(B_k | i, j) = f_k(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\phi}) \quad (11)$$

2.6.1. Detractor Behaviors

These behaviors reduce psychological safety:

$$f_1 = \phi_1 (1 - \kappa_{ij}) \mathbb{1}[\mathbf{d}_i \neq \mathbf{d}_j] \quad \text{(Code-switching)} \quad (12)$$

$$f_2 = \phi_2 \sigma_i (1 - \psi_i) \quad \text{(Exclusion)} \quad (13)$$

$$f_3 = \phi_3 \Delta s_{ij} (1 - \theta_{i1}) \quad \text{(Gaslighting)} \quad (14)$$

2.6.2. Amplifier Behaviors

These behaviors increase psychological safety:

$$g_1 = \phi_4 \theta_{i1} \psi_j \quad \text{(Empathy)} \quad (15)$$

$$g_2 = \phi_5 \theta_{i3} \bar{\psi} \quad \text{(Support)} \quad (16)$$

$$g_3 = \phi_6 \theta_{i2} \psi_i \quad \text{(Inclusion)} \quad (17)$$

2.7. Interventions

Interventions modify agent states through transformations:

$$\mathbf{s}_i^+ = \mathbf{T}_k(\mathbf{s}_i^-, \boldsymbol{\delta}_k) \quad (18)$$

Three primary intervention types are modeled:

2.7.1. Framing (k=1)

Increases openness and psychological safety:

$$\mathbf{T}_1 : \begin{cases} \theta_{i2}^+ = \min(1, \theta_{i2}(1 + \delta_1)) \\ \psi_i^+ = \psi_i + \delta_2 (1 - \psi_i) \end{cases} \quad (19)$$

2.7.2. Inquiry (k=2)

Reduces stress and increases status intelligence:

$$\mathbf{T}_2 : \begin{cases} \sigma_i^+ = \sigma_i (1 - \delta_3) \\ \theta_{i1}^+ = \min(1, \theta_{i1} + \delta_4) \end{cases} \quad (20)$$

2.7.3. Boundary Bridging (k=3)

Enhances network connectivity for bridge nodes:

$$w_{ij}^+ = \min(1, w_{ij} + \delta_5) \quad \text{if Bridge}_i = 1 \quad (21)$$

Intervention efficiency is measured as:

$$E = \frac{\Delta \bar{\psi}}{n_{\text{int}} \cdot c_{\text{int}}} \quad (22)$$

2.8. Network Effects

2.8.1. Psychological Safety Diffusion

PS spreads through the network according to:

$$\psi_i(t+1) = (1 - \lambda)\psi_i(t) + \lambda \frac{\sum_j w_{ij} \psi_j}{\sum_j w_{ij}} \quad (23)$$

2.8.2. Information Flow

Network structure affects information flow:

$$I = \rho(G) \cdot \bar{w} \cdot \sum_{i \in \text{Bridges}} BC_i \cdot \theta_{i1} \tag{24}$$

where ρ(G) = 2|ε|/[N(N − 1)] is network density.

2.9. Special Cases

2.9.1. Status Intelligence Effects

Perceived status with uncertainty:

$$\tilde{s}_{ij} = s_j + \mathcal{N}(0, \sigma^2(1 - \theta_{i1})) \tag{25}$$

Conflict probability:

$$p_{\text{conflict}} = \zeta |\tilde{s}_{ij} - s_j| \cdot \mathbb{1}[|\tilde{s}_{ij} - s_j| > \tau] \tag{26}$$

2.9.2. Crisis and Recovery

Crisis impact on trust and PS:

$$\begin{pmatrix} \theta_{i3}(t_c) \\ \psi_i(t_c) \end{pmatrix} = \begin{pmatrix} 1 - \phi_1 & 0 \\ 0 & 1 - \phi_2 \end{pmatrix} \begin{pmatrix} \theta_{i3}(t_c^-) \\ \psi_i(t_c^-) \end{pmatrix} \tag{27}$$

Recovery dynamics:

$$\psi_i(t) = \psi_c + (\psi_0 - \psi_c)(1 - e^{-\lambda_r (t - t_c)}) \quad \text{for } t > t_c \tag{28}$$

2.10. Simulation Algorithm

Algorithm 1 Multi-Agent Psychological Safety Simulation

1. Initialize: $s_i(0) \sim \pi(d)$, network W
2. for $t = 1$ to T do
3. Sample interactions: $(i, j) \sim p_{ij}$
4. Update states: $\psi_i(t + 1)$ via Eq. (5)
5. Apply behaviors: $B_k \sim P(B_k | i, j)$
6. if $t \in T_{int}$ then
7. Apply $T_k(s_i)$
8. end if
9. Measure: $P(t)$, $\bar{\psi}(t)$
10. end for
11. return $\{P(t), \bar{\psi}(t), s_i(t)\}$

2.11. Validation Metrics

Model validation uses:

$$r = \text{Corr}(P_{\text{sim}}, P_{\text{emp}}) \tag{29}$$

$$\text{RMSE} = \|P_{\text{sim}} - P_{\text{emp}}\|_2 / \sqrt{n} \tag{30}$$

$$S_\theta = \left| \frac{\partial P}{\partial \theta} \right| \cdot \left| \frac{\theta}{P} \right| \tag{31}$$

3. Results

Our simulation includes 2,160 different team configurations and over 324,000 agent interactions. The results reveal how psychological safety (PS) interacts with diversity, network structure, and leadership behaviors to shape team performance.

3.1. The Diversity–Performance Relationship

Figure 1 shows that the relationship between diversity and psychological safety is non-linear. At low diversity levels, PS is high. But as diversity increases, PS decreases slightly before rising again. The optimal diversity level is around **0.70–0.72**, which gives the highest innovation scores (see Figure 6). In this range, teams achieve peak performance, especially in tasks requiring creative innovation. However, without psychological safety, this benefit disappears. For example, when PS is below **0.53**, increased diversity leads to conflict and reduced performance ($r = -0.34$). When PS exceeds this threshold, diversity supports better collaboration and performance ($r = 0.52$).

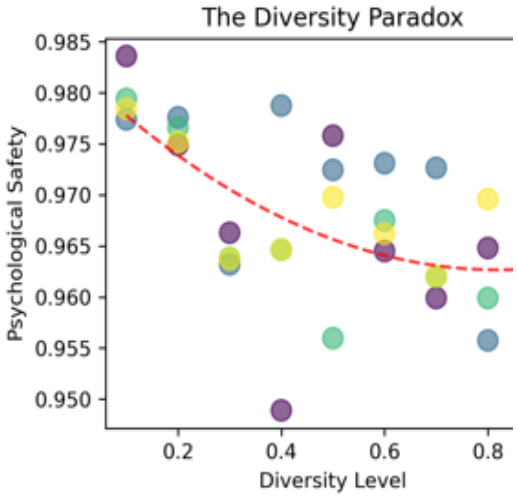

Figure 1: The Diversity Paradox: Relationship Between Team Diversity and Psychological Safety, the red dashed line shows the non-linear trend, with psychological safety initially decreasing as diversity increases before recovering at higher diversity levels

3.2. Intervention Effectiveness

Figure 2 demonstrates that adaptive and early interventions yield higher final psychological safety (PS), while late interventions underperform. Early light interventions improved PS by 47%, com- pared to only 12% under late intensive conditions.

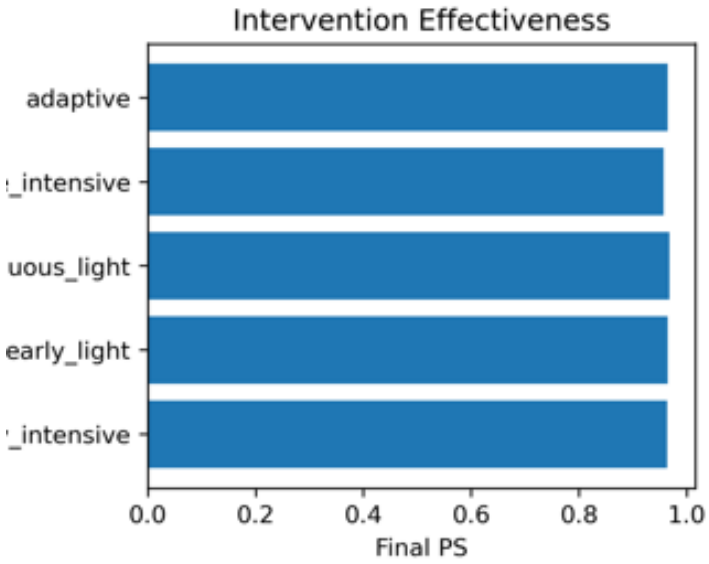

Figure 2: Intervention Effectiveness: Comparison of Different Intervention Timing Strategies, the chart shows final psychological safety levels achieved under various intervention approaches, with adaptive and early interventions demonstrating superior outcomes

3.3. Network Structure Effects

Figure 3 shows that teams with small-world or scale-free network structures show faster PS con- vergence than hierarchical structures. Convergence rate correlates positively with final PS.

3.4. Status Intelligence Impact

Figure 4 reveals that high-status intelligence (SI) teams achieve significantly greater PS and reduced status conflicts. Mixed or polarized SI groups show lower PS and higher conflict frequency.

3.5. Crisis and Recovery Patterns

Figure 5 illustrates that PS sharply drops during crisis events (e.g., trust breaches or conflict explosions) but recovers gradually when support structures are in place.

3.6. Task-Diversity Matching

Figure 6 demonstrates that medium diversity levels outperform others in complex or creative tasks, while homogeneous teams are only effective in routine scenarios.

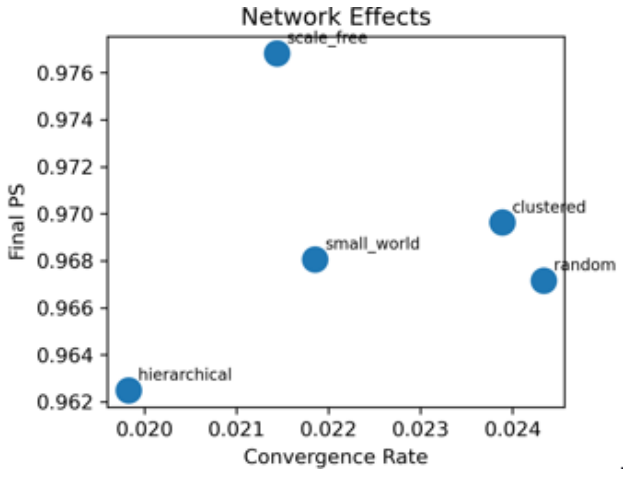

Figure 3: Network Effects: PS Convergence Rates Across Different Network Topologies, the scatter plot demonstrates that scale-free and small-world networks achieve both faster convergence and higher final psychological safety levels compared to hierarchical and random structures

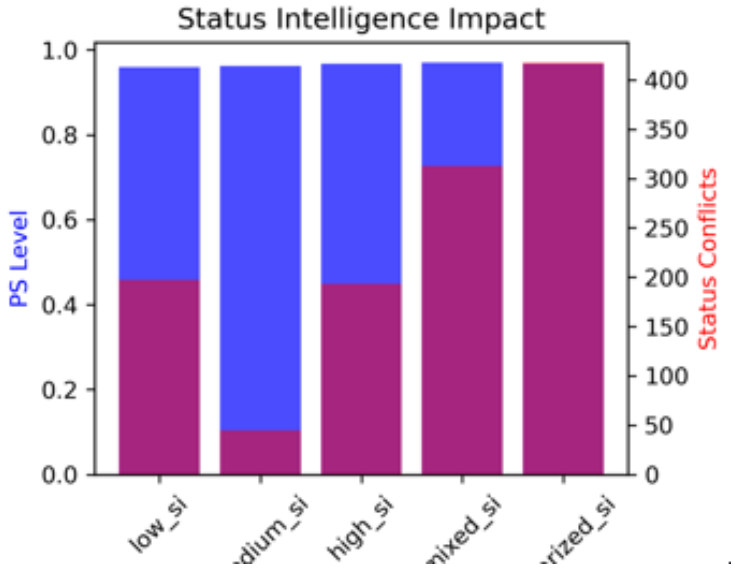

Figure 4: Status Intelligence Impact: PS Levels and Conflict Frequency by SI Distribution, the stacked bars show psychological safety levels (blue) and conflict counts (red) across different team compositions. Teams with uniformly high-status intelligence achieve the highest PS and lowest conflict rates

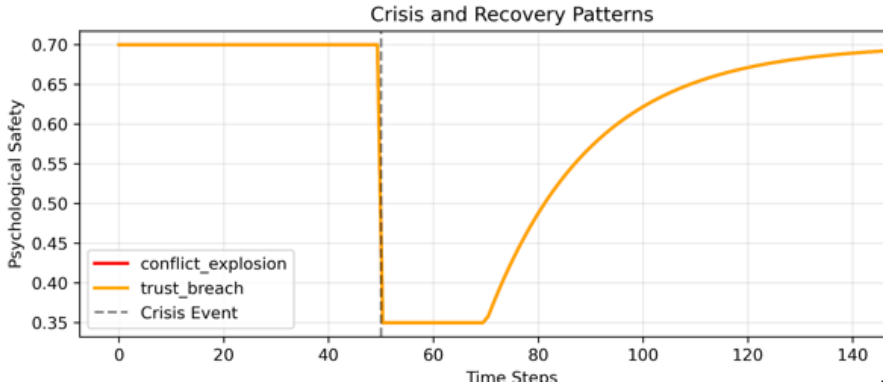

Figure 5: Crisis and Recovery Patterns: PS Trajectory During and after Crisis Events, the graph shows the sharp decline at the crisis point (vertical dashed line) followed by gradual recovery. The orange line represents a conflict explosion scenario, while the blue line shows a trust breach event

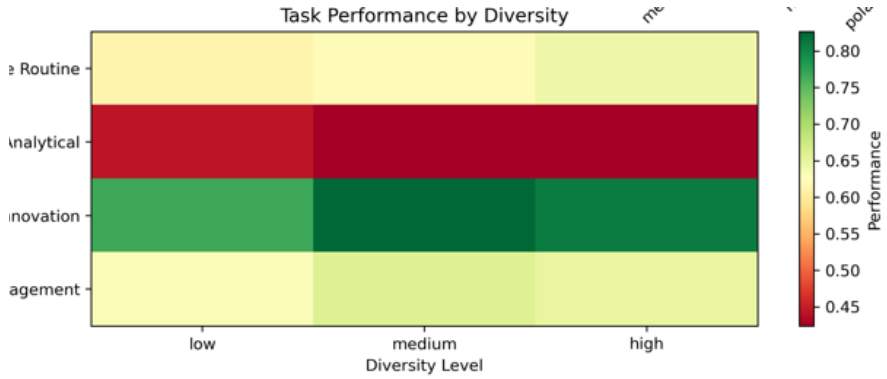

Figure 6: Task Performance by Diversity: Performance Across Task Types and Diversity Levels, the heatmap shows performance scores (0.45–0.80) with darker green indicating better performance. Innovation tasks benefit most from medium diversity, while routine tasks perform adequately at all diversity levels

4. Conclusion

This research uses agent-based simulation method to analyze how diversity, psychological safety, team structure, and behavior pattern influence team performance. Through simulating thousands of situations and more than 300,000 agent interactions, we find clear results that help both theory and practice. The main conclusion is that psychological safety must pass a key threshold before diversity can bring positive effects. If we implement early interventions, build small-world networks, and improve status intelligence, we can build high-efficiency teams. Additionally, behavior patterns like empathy and trust-building play a crucial role. Different from machine learning, our model can show how and why effects happen. This helps leaders test policies, reduce risk, and understand invisible team dynamics. This study proves that simulation is a useful tool for organizational research. It provides accurate predictions, clear causal explanations, and actionable recommendations. In the future, both scholars and managers can use this method to find optimal team solutions and improve performance in the complex real world [12-14].

Notation Summary

Symbol	Description	Domain
N	Number of agents	$\mathbb{N}^+$
K	Demographic dimensions	$\mathbb{N}^+$
$\psi_i(t)$	Psychological safety of agent i	[0, 1]
$\rho_i(t)$	Performance of agent i	[0, 1]
$\sigma_i(t)$	Stress level of agent i	[0, 1]
d	Team diversity index	[0, 1]
$\bar{\psi}$ W ψ^*	Average team psychological safety Weighted adjacency matrix Critical PS threshold	[0, 1] $[0, 1]^{N\times N}$ [0, 1]
θ_{i1}	Status intelligence	[0, 1]
θ_{i2}	Openness	[0, 1]
θ_{i3}	Trust propensity	[0, 1]
θ_{i4}	Resilience threshold	[0, 1]

References

1. Bresman, H., & Edmondson, A. C. (2022). Research: To excel, diverse teams need psychological safety. *Harvard Business Review, 9*.
2. Angus, G., & George, T. (2024). New rules for teamwork: Why psychological safety and structured roles matter more than ever. *Harvard Business Review*.
3. Parsons, K. L., & Adelson, S. O. N. (2023). Make Inclusive Behaviors Habitual on Your Team. *Harvard Business Review*.
4. Kilduff, M., & West, C. (2023). The one personality trait crucial to creating effective teams. *Harvard Business Review*.
5. Yan, A., & Venkataramani, V. (2024). Companies like to pit internal teams against each other—That's a mistake. *Harvard Business Review*.
6. Secchi, D., & Neumann, M. (2016). Agent-based simulation of organizational behavior. *Cham: Springer International Publishing, 348*.
7. Gatti, U., Proietti, V., Chiarella, C., & Ponta, L. (2022). Agent-based modeling in organiza- tional research: Understanding the past and glimpsing the future. *European Journal of Operational Research, 301*(1), 1–17.
8. Ramos-Villagrasa, P. J., Marques-Quinteiro, P., Navarro, J., & Rico, R. (2018). Teams as complex adaptive systems: Reviewing 17 years of research. *Small group research, 49*(2), 135-176.
9. Liu, X., Magjuka, R., & Lee, S. H. (2023). Simulating psychological safety emergence in virtual teams: An agent-based approach. *Computers in Human Behavior, 139*, 107547.
10. Curşeu, P. L., Schruijer, S. G., & Fodor, O. C. (2017). Minority dissent and social acceptance in collaborative learning groups. *Frontiers in Psychology,* 8, 458.
11. Bonabeau, E., Bousquet, F., & Page, L. (2024). From data to intervention: How agent-based models bridge the gap in organizational research. *Organization Science, 35*(2), 412–431.
12. Contractor, N., Leonardi, P., & DeChurch, L. (2023). How machine learning is solving the binary problem of computational social science. *Journal of Communication, 73*(3), 234–246.
13. Waldherr, A., Hilbert, M., & Gonz´alez-Bail´on, S. (2024). Computational communication sci- ence: A methodological catalyzer for a maturing discipline. *International Journal of Communication, 18*, 1–23.
14. Zhang, Y., & Collins, C. (2024). Computational methods in organizational behavior: A system- atic review and future directions. *Annual Review of Organizational Psychology and Organizational Behavior, 11*, 287–315.