International Journal of Education, Leadership, Artificial Intelligence, Computing,
DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

The Leadership Authenticity Dilemma: A Game-Theoretic Perspective on AI-Mediated Communication, Authenticity, Reciprocity and Interpersonal Trust

Bogdan COSTACHE
Bucharest University of Economic Studies, Bucharest, Romania

Abstract
The increasing integration of artificial intelligence (AI) into managerial communication is transforming the mechanisms through which leaders express authenticity, demonstrate relational commitment, and establish interpersonal trust. While AI-assisted communication offers substantial benefits in terms of efficiency, consistency, and linguistic quality, it also introduces strategic ambiguity regarding the intentionality, personal involvement, and genuine relational orientation underlying leadership messages. Consequently, employees may encounter increasing difficulties in distinguishing authentic expressions of leadership commitment from strategically constructed representations of authenticity.
*This study proposes the **Leadership Authenticity Signaling Game (LASG)**, a formal theoretical model integrating Signaling Game Theory and Trust Game Theory with authentic leadership, social exchange, and interpersonal trust perspectives. Employing a conceptual and analytical research methodology, the study conceptualizes leader-employee communication as a sequential strategic interaction under conditions of incomplete information and asymmetric knowledge. The proposed framework incorporates heterogeneous leader types, AI-mediated signal production costs, disclosure strategies, Bayesian belief updating, employee trust decisions, and subsequent reciprocal or opportunistic behavior.*
The central theoretical mechanism, termed the AI-Induced Authenticity Signal Dilution Effect (AISDE), examines how AI-assisted communication may reduce the differential costs associated with producing authentic-looking signals, potentially weakening their capacity to distinguish relationally committed leaders from strategically opportunistic actors. The analytical framework investigates the conditions under which AI mediation may facilitate transitions between separating, pooling, and semi-separating equilibria, thereby influencing signal credibility, perceived leadership authenticity, and interpersonal trust. Particular attention is devoted to the potential role of verifiable relational commitment, authorship transparency, and behavioral accountability in preserving informative signaling and sustaining cooperative interactions.
The study seeks to extend authentic leadership research beyond predominantly psychological and behavioral explanations by introducing a game-theoretic perspective on authenticity as a strategically communicated and relationally interpreted phenomenon. Rather than conceptualizing AI as inherently detrimental to leadership authenticity, the proposed model identifies the strategic conditions under which technological mediation may either undermine or preserve relational credibility. The framework provides a theoretical foundation for responsible AI integration, trust-sensitive managerial communication, and leadership development in increasingly AI-mediated organizational environments, while establishing testable propositions for subsequent empirical research.

Keywords: Authentic Leadership; Leadership Authenticity Signaling Game (LASG); AI-Induced Authenticity Signal Dilution Effect (AISDE); Signaling Games; Trust Games; Bayesian Games; Perfect Bayesian Equilibrium.

1. INTRODUCTION

The rapid integration of artificial intelligence (AI) into organizational processes is fundamentally reshaping managerial practices, interpersonal communication, and the exercise of leadership. Beyond supporting analytical and operational activities, generative AI increasingly participates in the production of messages through which leaders communicate expectations, provide feedback, express empathy, and demonstrate relational commitment. AI-mediated communication represents an emerging paradigm in which intelligent systems modify, augment, or generate interpersonal messages on behalf of human communicators [1]. Although this transformation creates opportunities for

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

managerial efficiency and communication effectiveness, it also introduces challenges concerning human agency, accountability, and the relational foundations of organizational trust. These challenges reflect the broader automation–augmentation paradox, whereby the benefits of AI depend partly on how technological capabilities interact with human judgment and responsibility [2].

Authenticity constitutes a significant yet contested dimension of contemporary leadership research. Authentic leadership has traditionally been associated with self-awareness, relational transparency, ethical conduct, and consistency between personal values and observable behavior [3]. However, critical perspectives challenge the assumption that authenticity represents an inherently positive, stable, and universally desirable leadership characteristic [4]. Alternative approaches emphasize its relational and socially constructed nature, suggesting that authenticity emerges through interactions between individuals and their organizational environments [5]. Furthermore, the authenticity paradox highlights the tension between remaining consistent with one's established identity and adapting behavior to changing professional demands [6]. These perspectives suggest that AI-assisted communication should not automatically be interpreted as evidence of inauthentic leadership.

Nevertheless, AI introduces a distinctive problem concerning the credibility of authenticity-related signals. Organizational trust involves the willingness of individuals to accept vulnerability based on positive expectations regarding another party's ability, benevolence, and integrity [7]. Within leader–employee relationships, such expectations are partly shaped by communicative behavior and perceived relational investment. Recent research indicates that AI-mediated communication can influence both linguistic performance and interpersonal evaluations [8]. Moreover, individuals may experience supportive AI-generated communication positively while evaluating the same interaction less favorably when its artificial origin is disclosed [9]. Research also suggests that human judgments regarding AI-generated language are frequently unreliable [10], while disclosure of AI involvement may, under certain conditions, produce a trust penalty [11].

These developments raise a strategic question that extends beyond conventional psychological explanations of leadership authenticity. If generative AI enables leaders to produce persuasive expressions of empathy, commitment, and personal involvement at substantially reduced costs, employees may encounter increasing difficulties in determining whether such expressions reflect genuine relational intentions. Signaling theory provides an appropriate analytical foundation for examining this problem because it explains how observable signals convey otherwise unobservable characteristics under conditions of information asymmetry [12,13]. Similarly, the Trust Game provides a formal framework for investigating vulnerability, reciprocity, and the possibility of opportunistic behavior within interpersonal exchanges [14].

Although existing research provides valuable insights into authentic leadership, organizational trust, and AI-mediated communication, the strategic interdependence between leaders' communication choices and employees' trust decisions remains comparatively underexplored. Emerging research examining leaders' AI use and disclosure brings these issues closer to the leadership domain [15]. However, a complementary analytical perspective is needed to explain how AI modifies the incentive structures governing authenticity-related signaling, how employees update their beliefs regarding leaders' relational intentions, and under what conditions communication signals retain or lose their informational value.

To address this research gap, the present study proposes the **Leadership Authenticity Signaling Game (LASG)**, a formal game-theoretic framework integrating Signaling Games and Trust Games within a sequential interaction characterized by incomplete information. The model distinguishes between underlying relational orientation, communicated authenticity, perceived authenticity, and behavioral reciprocity. Its central mechanism, termed the **AI-Induced Authenticity Signal Dilution Effect (AISDE)**, examines whether AI-mediated communication reduces the differential costs of producing authenticity-related signals, potentially allowing opportunistically oriented leaders to imitate signals previously associated with genuine relational commitment.
Accordingly, the study addresses the following central research question:

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Under what strategic conditions does AI-mediated leadership communication weaken or preserve the informational credibility of authenticity-related signals and the formation of interpersonal trust?

The research pursues four interconnected objectives: developing a formal representation of AI-mediated leader–employee interactions; identifying the mechanisms through which AI modifies signaling costs and employee beliefs; deriving the conditions supporting separating, pooling, and semi-separating equilibria; and examining whether verifiable relational commitment, transparency, and behavioral accountability can preserve or restore signal credibility.

The intended theoretical contribution is threefold. First, the study extends authentic leadership research by conceptualizing communicated authenticity as a strategic signal rather than exclusively as an individual leadership characteristic. Second, it extends the application of Signaling Games and Trust Games to AI-mediated organizational relationships by introducing technological mediation as a parameter influencing signaling costs and strategic incentives. Third, it proposes AISDE as a theoretically testable mechanism connecting technological efficiency with potential changes in relational credibility and interpersonal trust.

Rather than treating AI adoption as inherently incompatible with authentic leadership, the proposed framework seeks to identify the conditions under which technological mediation supports or undermines sustainable leader–employee relationships. Its practical relevance extends to organizational management, responsible AI integration, educational leadership, and the development of leadership competencies in increasingly AI-mediated professional environments.

2. LITERATURE REVIEW

The theoretical development of the Leadership Authenticity Signaling Game (LASG) requires the integration of three complementary research domains: the relational foundations of authentic leadership and interpersonal trust, the strategic mechanisms underlying signaling and reciprocity, and the technological transformation of interpersonal communication through artificial intelligence. Although these domains have developed largely through distinct theoretical traditions, their integration provides a foundation for examining leadership authenticity as both a relational phenomenon and a strategic interaction under conditions of incomplete information.

2.1. Authentic Leadership as a Relational and Contextual Phenomenon

Authentic leadership research has traditionally emphasized self-awareness, relational transparency, balanced information processing, and consistency between internal values and observable behavior [3]. Nevertheless, the conceptualization of authenticity as a relatively stable and inherently desirable leadership characteristic has attracted substantial criticism. Alvesson and Einola [4] challenge the excessive positivity underlying authentic leadership research, while Gardner et al. [16] acknowledge continuing theoretical debates concerning its conceptual boundaries, explanatory value, and empirical operationalization. Furthermore, evidence of substantial conceptual overlap between authentic and transformational leadership raises questions regarding the distinctiveness of authenticity as an independent explanatory construct [17].

Alternative perspectives conceptualize authenticity as a relational and socially negotiated phenomenon rather than an exclusively individual characteristic. Algera and Lips-Wiersma [5] emphasize the organizational conditions that enable individuals to express authenticity, while the life-stories approach developed by Shamir and Eilam-Shamir [18] connects leadership authenticity with narrative identity and the interpretation of personal experience. Ibarra [6] further demonstrates the conceptual tension between identity consistency and behavioral adaptation, suggesting that leadership development may require individuals to move beyond established patterns of self-expression.

The contextual character of authenticity is particularly relevant to contemporary organizational environments. Liu et al. [19] examine the gendered construction of authentic leadership, while Ayaz et al. [20] explore the constraints experienced by atypical leaders whose authentic self-expression may conflict with prevailing organizational expectations. These contributions suggest that perceived authenticity depends not only on leaders' underlying intentions but also on the interpretative frameworks through which followers evaluate their behavior.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Within AI-mediated communication, this distinction becomes particularly important. A leader may employ AI to improve the expression of genuine intentions, whereas another may use similar technological capabilities to simulate relational commitment. Consequently, LASG distinguishes between underlying relational orientation, communicated authenticity, and perceived authenticity, treating their potential divergence as a strategic and informational problem.

2.2. Interpersonal Trust, Social Exchange, and Reciprocity

Interpersonal trust represents a fundamental mechanism through which organizational relationships develop and cooperative behavior becomes sustainable. Mayer et al. [7] conceptualize trust as the willingness to accept vulnerability based on expectations regarding another party's ability, benevolence, and integrity. Within leadership relationships, these expectations are influenced by both observable managerial behavior and employees' interpretations of leaders' underlying intentions.

Social Exchange Theory provides a complementary explanation of trust formation by emphasizing reciprocal obligations and the relational consequences of interpersonal exchanges. Cropanzano et al. [21] highlight the complexity of social exchange mechanisms, particularly the relationship between exchange processes, reciprocal behavior, and organizational outcomes. Similarly, Colquitt et al. [22] examine the role of trust in reducing uncertainty and strengthening social exchange relationships. The association between authentic leadership, trust, and social exchange has also been explored in contemporary organizational research [23].

From a game-theoretic perspective, the Trust Game introduced by Berg et al. [14] formalizes the relationship between vulnerability and reciprocity through a sequential interaction in which one actor voluntarily transfers resources to another, who subsequently decides whether to reciprocate. However, experimental evidence indicates considerable heterogeneity in reciprocal behavior, suggesting that individuals respond differently to opportunities for cooperation and personal gain [24].

Within LASG, employees' trust decisions are conceptualized as strategic responses to uncertainty regarding leaders' future behavior. Unlike conventional formulations of the Trust Game, the proposed framework introduces a preliminary signaling stage during which leaders communicate information that may influence employees' expectations of reciprocity. Trust is therefore modeled not simply as a consequence of perceived authenticity but as a decision informed by beliefs, anticipated benefits, and potential relational losses.

2.3. Signaling Games, Information Asymmetry, and Strategic Communication

Signaling Theory explains how actors communicate privately held characteristics through observable actions under conditions of information asymmetry. Spence's foundational contribution [12] demonstrates how differences in signaling costs may enable receivers to distinguish between heterogeneous sender types. Subsequent organizational research has extended signaling theory to examine signal credibility, information asymmetry, sender characteristics, and receiver interpretation across managerial contexts [13].

Game-theoretic approaches provide a formal analytical framework for examining these mechanisms. Sequential games with incomplete information enable the representation of strategic interactions in which actors make decisions based on beliefs about privately held characteristics and anticipated responses [25]. Within such frameworks, separating equilibria may emerge when different sender types select distinguishable signals, whereas pooling equilibria occur when heterogeneous senders adopt observationally equivalent strategies. Research on strategic communication under partial information further highlights the role of trust and risk tolerance in determining how communicated signals are interpreted [26].

Applied to leadership, expressions of empathy, personal attention, and relational commitment may function as signals of underlying managerial intentions. Their informational value, however, depends on whether different types of leaders face sufficiently distinct incentives or costs when producing them. If technological assistance reduces the cost of imitating authenticity-related signals, the compatibility between communicated behavior and underlying relational orientation may become increasingly uncertain.

LASG extends this reasoning by introducing AI mediation as a factor that may alter both signaling costs and the incentive compatibility conditions supporting informative communication.

2.4. Artificial Intelligence and the Transformation of Authenticity Signals

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

AI-mediated communication introduces new opportunities for individuals to improve the clarity, emotional sensitivity, and persuasive effectiveness of interpersonal messages [1,8]. Nevertheless, the relational consequences of such mediation depend not only on communication quality but also on how recipients interpret human involvement and technological agency.

Research on human trust in AI emphasizes the importance of contextual and technological characteristics in shaping trust judgments [27]. However, trust in an AI system should be distinguished from interpersonal trust in a human leader who employs that system. LASG primarily addresses the latter, examining AI as a technological intermediary that modifies the conditions under which relational intentions are communicated and interpreted.

Empirical research has demonstrated limitations in individuals' ability to distinguish human-generated from AI-generated language [10]. Furthermore, positive evaluations of supportive communication may diminish when recipients become aware of its artificial origin [9]. Studies examining AI disclosure similarly suggest that transparency regarding technological involvement may produce unintended interpersonal trust penalties [11].

Related evidence from marketing communication indicates that AI authorship can influence perceived authenticity and emotional evaluations [28]. Although findings from consumer research cannot be directly generalized to leader–employee relationships, they provide complementary insights into how attributed authorship shapes message interpretation. Research by Dorigoni and Giardino [29] further indicates that perceived emotional authenticity depends on the relational identity and attributed involvement of the message producer.

Emerging leadership research explicitly addresses the consequences of leaders' AI use and disclosure for employee responses [15]. Nevertheless, an important theoretical distinction remains between the psychological consequences of AI disclosure and the strategic conditions that determine whether authenticity-related signals remain informative.

The proposed LASG framework addresses this distinction by conceptualizing AI mediation as a technological modification of the signaling environment rather than an inherently negative determinant of leadership authenticity.

2.5. Theoretical Integration and Research Gap

The reviewed literature provides complementary explanations of authenticity, interpersonal trust, strategic communication, and technological mediation. Authentic leadership research establishes the relational and contextual significance of authenticity, while Social Exchange Theory explains the role of reciprocity in organizational relationships. Signaling Games provide the mathematical foundations for examining privately held characteristics and signal credibility, whereas Trust Games formalize vulnerability and subsequent reciprocal or opportunistic behavior.

However, the integration of these theoretical perspectives within a unified formal model of AI-mediated leadership communication remains a research opportunity. Specifically, this study investigates whether AI-induced reductions in the differential costs of authenticity-related signaling can alter the strategic conditions supporting informative communication and interpersonal trust.

The Leadership Authenticity Signaling Game consequently integrates a signaling stage with a subsequent trust and reciprocity stage, allowing the analytical examination of the proposed AI-Induced Authenticity Signal Dilution Effect. This integration shifts the emphasis from whether AI use necessarily undermines authenticity toward identifying the conditions under which technological mediation weakens, preserves, or potentially restores the informational credibility of leadership communication.

Table 1. Theoretical Foundations and Conceptual Contributions to the Leadership Authenticity Signaling Game (LASG)

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Theoretical foundation	Central theoretical construct	Contribution to LASG
Authentic Leadership Theory	Authenticity and relational transparency	Distinguishing underlying relational orientation from communicated authenticity
Organizational Trust Theory	Trustworthiness and vulnerability	Defining employees' trust decisions and relational exposure
Social Exchange Theory	Reciprocity and relational obligations	Explaining reciprocal and opportunistic behavior
Signaling Theory	Signal credibility and information asymmetry	Establishing the relationship between leader types, signaling costs, and observable messages
Signaling Games	Strategic incentives and Bayesian beliefs	Deriving separating, pooling, and potentially semi-separating equilibria
Trust Games	Vulnerability and strategic reciprocity	Formalizing employee trust decisions and subsequent leadership responses
Automation–Augmentation Paradox	Technological efficiency and human agency	Conceptualizing AI as a modifier of communication and signaling costs
AI-Mediated Communication	Technological mediation and authorship	Explaining changes in message production and perceived human involvement
Human–AI Trust and Disclosure Research	Transparency, attribution, and trust	Informing assumptions regarding technological disclosure and signal interpretation
LASG – Proposed theoretical integration	AI-Induced Authenticity Signal Dilution Effect (AISDE)	Examining the conditions under which AI changes signal informativeness, strategic equilibria, and interpersonal trust

Source: Authors' conceptual synthesis based on the reviewed literature.

3. METHODOLOGY

3.1. Research Design and Analytical Approach

This study adopts a conceptual and formal analytical research design aimed at developing the Leadership Authenticity Signaling Game (LASG) as a theoretical representation of strategic interactions between leaders and employees in AI-mediated organizational environments. The methodological approach combines interdisciplinary theoretical integration, deductive mathematical modeling, and equilibrium analysis. Rather than estimating empirical relationships between predefined variables, the study examines the logical consequences of explicitly formulated behavioral assumptions and strategic incentive structures.

The research process comprises four interconnected stages. The first stage involves the conceptual integration of Authentic Leadership Theory, Social Exchange Theory, organizational trust research, Signaling Games, and Trust Games. The second stage translates the identified theoretical constructs into formal game-theoretic components, including players, privately held characteristics, available strategies, information structures, and utility functions. The third stage derives the conditions supporting alternative strategic equilibria and examines how AI mediation modifies these conditions. The fourth stage interprets the analytical results through theoretically testable propositions and examines their implications for leadership authenticity, interpersonal trust, and responsible AI adoption.

The methodological framework follows a theory-building approach in which behavioral assumptions are informed by the reviewed literature, while mathematical results are derived deductively.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Accordingly, the proposed model constitutes an analytical representation of strategic behavior rather than an empirically calibrated description of organizational reality.

3.2. Game Structure and Information Asymmetry

LASG is conceptualized as a sequential Bayesian game involving two strategically interdependent players: a leader L and an employee E. The leader possesses private information regarding their underlying relational orientation, while the employee observes communication signals and forms expectations concerning the leader's subsequent behavior.

The set of possible leader types is defined as:

$$\Theta = \{\theta_R, \theta_O\} \tag{1}$$

where θ_R denotes a relationally oriented leader and θ_O denotes an opportunistically oriented leader. These categories represent stylized analytical types rather than exhaustive psychological classifications.

The employee's initial belief regarding the leader's type is represented by a common prior probability:

$$P(\theta = \theta_R) = p,\ P(\theta = \theta_O) = 1 - p \tag{2}$$

where $0 < p < 1$.

The leader observes their own type before selecting a communication strategy. The employee does not directly observe this characteristic but receives an observable signal and subsequently decides whether to extend interpersonal trust.

AI is introduced as a technological mediation parameter rather than an independent strategic player:

$$\alpha \in [0,\ 1] \tag{3}$$

where $\alpha = 0$ represents communication without AI assistance, while $\alpha = 1$ represents the maximum degree of AI mediation considered within the model. Intermediate values represent different levels of technological involvement.

In the baseline specification, α is treated as an exogenous, commonly known parameter, allowing the analytical isolation of its effects on signaling costs. An extended specification subsequently considers disclosure and uncertainty regarding actual AI involvement.

3.3. Strategic Choices and Sequence of Interaction

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

The leader's communication strategy is represented by a binary signaling space:

$$S_L = \{s_H, s_L\} \tag{4}$$

where s_H represents a communication signal expressing a high degree of apparent relational commitment, while s_L represents a conventional communication signal conveying limited relational investment.

Importantly, s_H does not necessarily indicate genuine authenticity. Both leader types may produce this signal, although they may face different production costs and behavioral incentives.

Following the observation of the communication signal, the employee selects a trust-related action:

$$T_E = \{0, 1\} \tag{5}$$

where t = 1 denotes the decision to extend trust and accept relational vulnerability, while t = 0 denotes the decision to withhold trust.

Conditional upon the employee extending trust, the leader selects a subsequent behavioral response:

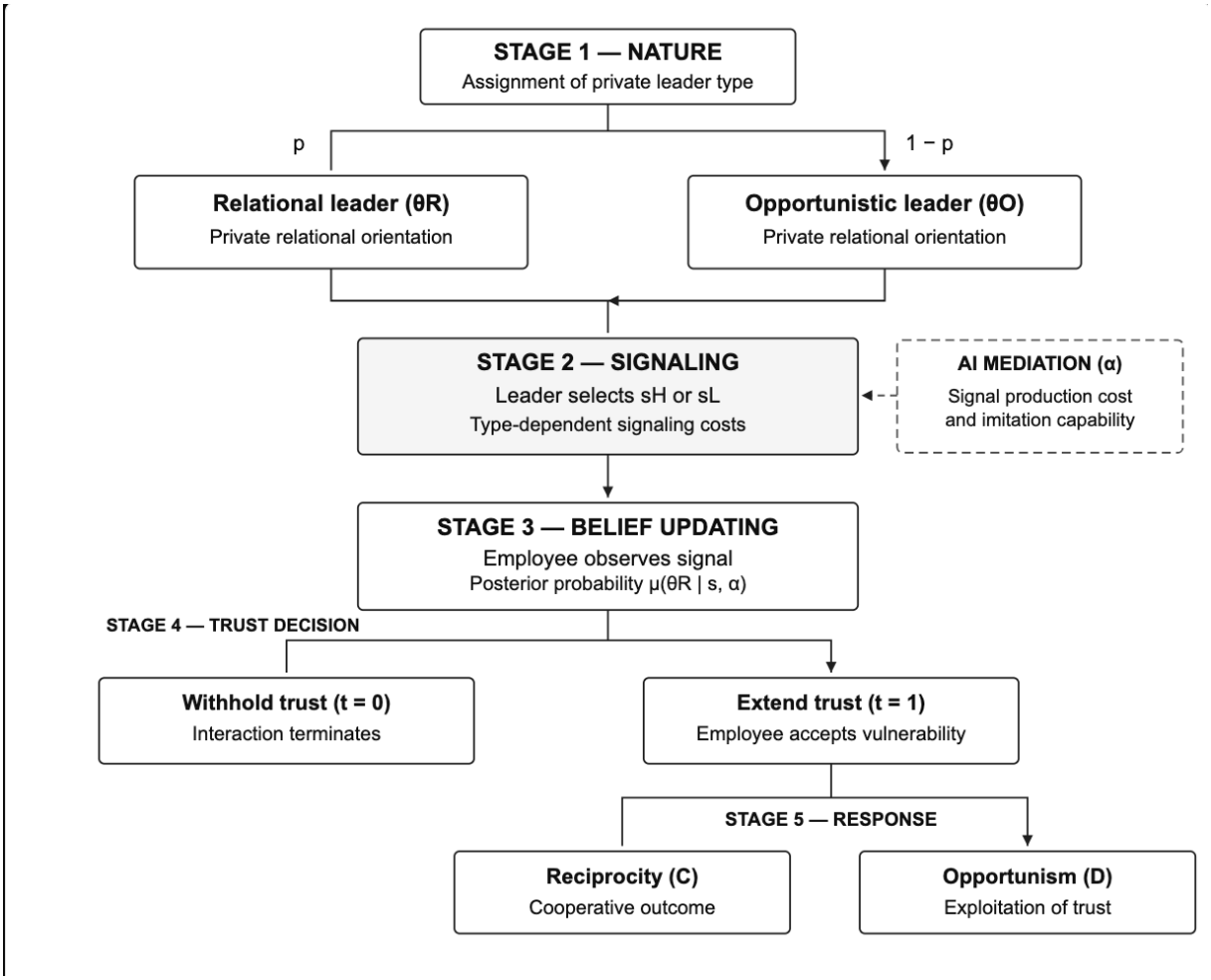
$$R_L = \{C, D\} \tag{6}$$

where C represents reciprocal behavior consistent with relational commitment, while D represents opportunistic behavior that exploits the vulnerability accepted by the employee.

The resulting game integrates a preliminary signaling interaction with a subsequent trust and reciprocity interaction. This sequential architecture allows communicated authenticity to be distinguished analytically from demonstrated relational trustworthiness.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Figure 1. Sequential Architecture of the Leadership Authenticity Signaling Game (LASG)



Source: Authors' theoretical development. The figure represents the sequential architecture of LASG rather than its complete extensive-form representation.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

3.4. Modeling Assumptions and AI-Induced Signal Dilution

The analytical development of LASG is based on six modeling assumptions. These assumptions establish the conditions under which strategic equilibria can be derived without presupposing that AI mediation necessarily undermines interpersonal trust.

Assumption 1 (A1) – Leader heterogeneity.
Leaders differ in their underlying preferences regarding relational commitment, reciprocity, and opportunistic behavior.

Assumption 2 (A2) – Information asymmetry.
Employees cannot directly observe leaders' underlying relational orientation before making trust-related decisions.

Assumption 3 (A3) – Differential signaling costs.
Producing a high-authenticity signal imposes a higher net cost on opportunistically oriented leaders than on relationally oriented leaders in the absence of AI mediation.

Let the incremental cost of producing a high-authenticity signal be:

$$c_i(\alpha) = c(s_H, \theta_i, \alpha) - c(s_L, \theta_i, \alpha) \tag{7}$$

where $i \in \{R, O\}$. The baseline cost differential is therefore:

$$c_O(0) > c_R(0) \geq 0 \tag{8}$$

Assumption 4 (A4) – AI-mediated cost reduction.
AI assistance may reduce the incremental cost of producing authenticity-related communication signals.

$$\frac{\partial c_i(\alpha)}{\partial \alpha} \leq 0,\ i \in \{R, O\} \tag{9}$$

Assumption 5 (A5) – Differential technological impact.
AI mediation may reduce signaling costs more substantially for opportunistically oriented leaders than for relationally oriented leaders.

Defining the difference between type-dependent signaling costs as:

$$\Delta c(\alpha) = c_O(\alpha) - c_R(\alpha) \tag{10}$$

the signal dilution assumption becomes:

$$\frac{d\Delta c(\alpha)}{d\alpha} < 0 \tag{11}$$

This assumption establishes the technological cost-compression mechanism underlying the proposed AI-Induced Authenticity Signal Dilution Effect (AISDE). Whether such compression actually reduces equilibrium signal informativeness will be determined through the subsequent analytical development.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Assumption 6 (A6) – Preference-based reciprocity.

Leaders differ in the utility associated with reciprocal and opportunistic responses. These preferences, together with the benefits and costs associated with each behavioral response, determine their continuation strategies through utility maximization.

The final assumption ensures that reciprocal or opportunistic behavior is derived from strategic incentives rather than imposed solely through the classification of leader types.

3.5. Bayesian Belief Updating and Utility Specification

Let $\sigma_i(s \mid \alpha)$ represent the probability that a leader of type i selects signal s, given the degree of AI mediation.

Following signal observation, the employee updates the prior probability of encountering a relationally oriented leader according to Bayes' rule:

$$\mu(s, \alpha) = \frac{p\sigma_R(s \mid \alpha)}{p\sigma_R(s \mid \alpha) + (1-p)\sigma_O(s \mid \alpha)} \tag{12}$$

whenever the denominator is strictly positive.

The employee's posterior belief represents the perceived probability that the leader possesses a relational orientation. It is therefore distinguished from the probability of actual reciprocal behavior.

Let q_R and q_O denote the equilibrium probabilities of reciprocity associated with the two leader types, conditional on the observed signal and the employee extending trust. The expected probability of reciprocity becomes:

$$\rho = \mu q_R + (1-\mu) q_O \tag{13}$$

The employee's expected utility from extending trust is specified as:

$$\mathbb{E}[U_E \mid t = 1] = \rho B - (1-\rho)K \tag{14}$$

where $B > 0$ represents the benefit associated with reciprocal behavior and $K > 0$ represents the loss resulting from opportunistic behavior. The utility of withholding trust is normalized to zero.

The leader's utility function incorporates the benefits associated with employee trust, the utility consequences of subsequent behavioral responses, and the costs of signal production:

$$U_L(\theta_i, s, t, r; \alpha) = tV_i(r) - c(s, \theta_i, \alpha) \tag{15}$$

where $V_i(r)$ denotes the type-dependent continuation payoff associated with behavioral response r.

These utility specifications provide the analytical foundation for deriving the employee's trust threshold, leader incentive compatibility constraints, and equilibrium signaling strategies.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

3.6. Equilibrium Analysis and Theoretical Validation

The principal solution concept adopted for LASG is Perfect Bayesian Equilibrium (PBE), which combines sequential rationality with Bayesian consistency of beliefs. An equilibrium consists of leader signaling strategies, employee trust strategies, leader continuation strategies, and posterior beliefs.

The analytical procedure employs backward induction to determine optimal continuation behavior, followed by Bayesian belief updating and the derivation of incentive compatibility conditions.

Three principal equilibrium configurations are investigated: separating equilibria, in which different leader types select distinguishable signals; pooling equilibria, in which both types select the same signal; and semi-separating equilibria, in which at least one leader type adopts a mixed signaling strategy while signals retain partial informational value.

Comparative statics are subsequently employed to examine how changes in AI mediation, prior beliefs, signaling costs, relational benefits, and vulnerability losses affect equilibrium feasibility and employee trust decisions. Analytical extensions consider the potential effects of verifiable relational commitment, disclosure, and behavioral accountability on the strategic credibility of authenticity-related communication.

Theoretical validation relies on mathematical consistency, sequential rationality, the examination of boundary conditions, and sensitivity analysis using illustrative parameter configurations. No primary empirical data are collected. Consequently, numerical examples represent hypothetical analytical scenarios rather than observed organizational behavior, and the proposed theoretical relationships remain subject to subsequent empirical testing.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Table 2. Formal Components and Parameters of the LASG Model

Component	Notation	Definition
Leader	L	Informed strategic player
Employee	E	Uninformed strategic player
Relational leader type	θ_r	Leader with relationally oriented preferences
Opportunistic leader type	θ_o	Leader with opportunistically oriented preferences
Prior probability	p	Initial probability attributed to the relational leader type
High-authenticity signal	s_h	Communication expressing apparent relational commitment
Conventional signal	s_l	Communication expressing limited relational investment
AI mediation	α	Degree of technological involvement
Incremental signaling costs	$c_i(\alpha)$	Additional cost of producing a high-authenticity signal
Signaling cost differential	$\Delta c(\alpha)$	Difference between type-dependent incremental signaling costs
Signaling strategy	$\sigma_i(s \mid \alpha)$	Probability of selecting a particular signal
Posterior belief	μ	Updated probability of relational leader orientation
Trust decision	t	Decision to extend or withhold trust
Reciprocal behavior	C	Cooperative behavioral response
Opportunistic behavior	D	Exploitative behavioral response
Type-specific reciprocity	q_i	Equilibrium probability of reciprocal behavior
Expected reciprocity	ρ	Employee's expected probability of reciprocal behavior
Relational benefit	B	Employee's benefit from successful reciprocity
Vulnerability loss	K	Employee's loss resulting from opportunism
Leader continuation payoff	$V_i(r)$	Type-dependent utility associated with behavioral response
Utility functions	U_l, U_e	Strategic payoffs of the two players
Equilibrium concept	PBE	Perfect Bayesian Equilibrium

Source: Authors' formal specification of the LASG model.

4. RESULTS – FORMAL DEVELOPMENT OF THE LEADERSHIP AUTHENTICITY SIGNALING GAME (LASG)

The analytical development of LASG establishes the strategic conditions under which AI-mediated communication may influence authenticity-related signaling, employee trust decisions, and subsequent relational outcomes. Building upon the methodological framework and assumptions A1–A6, this section derives the equilibrium conditions governing leader–employee interactions and formally examines the proposed AI-Induced Authenticity Signal Dilution Effect (AISDE).

The analysis proceeds from optimal continuation behavior to employee trust decisions, leader signaling incentives, equilibrium configurations, and comparative statics. All mathematical results are conditional upon the specified assumptions and utility functions.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

4.1. Optimal Continuation Behavior and the Employee Trust Threshold

Following the employee's decision to extend trust, the leader selects the behavioral response that maximizes their continuation utility:

$$r_i^* = \underset{r \in \{C, D\}}{arg\ max}\ V_i(r) \tag{16}$$

To obtain a tractable baseline specification, the following preference conditions are introduced:

$$V_R(C) > V_R(D) \tag{17}$$

$$V_O(D) > V_O(C) \tag{18}$$

Under these conditions, relationally oriented leaders reciprocate employee trust, whereas opportunistically oriented leaders exploit the vulnerability associated with that trust.

Consequently:

$$q_R = 1,\ q_O = 0 \tag{19}$$

Substituting Equation (19) into Equation (13) yields:

$$\rho = \mu \tag{20}$$

This equality represents a consequence of the baseline preference specification rather than a general equivalence between perceived authenticity and actual reciprocity.

The employee's expected utility from extending trust becomes:

$$\mathbb{E}[U_E \mid t = 1] = \mu B - (1 - \mu)K \tag{21}$$

Trust is strategically rational whenever:

$$\mu B - (1 - \mu)K \geq 0 \tag{22}$$

Rearranging Equation (22) produces the critical trust threshold:

$$\boxed{\mu^* = \frac{K}{B + K}} \tag{23}$$

The employee's optimal trust strategy is therefore:

$$t^*(\mu) = \begin{cases} 1, & \mu > \mu^* \\ 0, & \mu < \mu^* \\ \text{either action or mixing}, & \mu = \mu^* \end{cases} \tag{24}$$

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

The threshold increases with the potential loss associated with opportunistic behavior and decreases with the expected benefit of reciprocity:

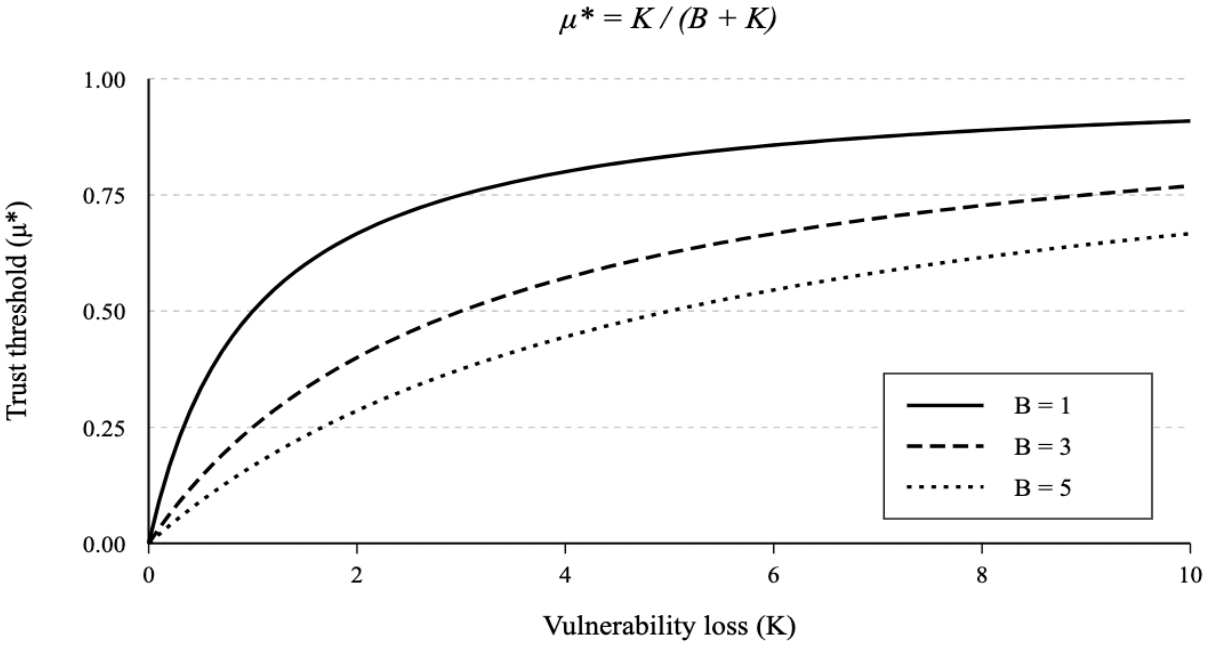
$$\frac{\partial \mu^*}{\partial K} = \frac{B}{(B+K)^2} > 0 \quad (25)$$

$$\frac{\partial \mu^*}{\partial B} = -\frac{K}{(B+K)^2} < 0 \quad (26)$$

Proposition 1 (P1) – Vulnerability-dependent trust threshold.

Within LASG, employees require stronger posterior beliefs regarding leaders' relational orientation when the potential consequences of opportunistic behavior increase. Conversely, greater anticipated benefits from reciprocity reduce the minimum belief required to extend trust.

Figure 2. Relationship between Vulnerability Loss and the Employee Trust Threshold



Source: Authors' analytical representation based on Equation (23).
Parameter values are illustrative.

4.2. Leader Incentives and Separating Equilibrium

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Let the optimal continuation payoffs associated with obtaining employee trust be:

$$W_R = V_R(C) > 0$$
$$W_O = V_O(D) > 0 \quad (27)$$

Normalizing the cost of the conventional signal to zero for each leader type, the reduced-form leader utility becomes:

$$U_i(s, t; \alpha) = tW_i - \mathbb{1}_{\{s = sH\}} c_i(\alpha) \quad (28)$$

where the indicator function equals one when the leader selects the high-authenticity signal and zero otherwise.

A separating equilibrium occurs when relationally oriented leaders select the high-authenticity signal and opportunistically oriented leaders select the conventional signal:

$$\sigma_R(s_H) = 1, \sigma_O(s_H) = 0 \quad (29)$$

Bayesian updating subsequently yields:

$$\mu(s_H) = 1, \mu(s_L) = 0 \quad (30)$$

Since the employee trust threshold lies strictly between zero and one, the corresponding employee strategy is to trust after observing s_H and withhold trust after observing s_L.

For the relational leader, the incentive compatibility condition is:

$$W_R - c_R(\alpha) \geq 0 \quad (31)$$

For the opportunistic leader, the incentive compatibility condition is:

$$0 \geq W_O - c_O(\alpha) \quad (32)$$

Combining these conditions establishes the feasibility of the separating equilibrium:

$$c_R(\alpha) \leq W_R, c_O(\alpha) \geq W_O \quad (33)$$

Proposition 2 (P2) – Incentive-compatible authenticity signaling.

Authenticity-related communication constitutes a credible separating signal when its incremental production cost remains sufficiently low for relationally oriented leaders and sufficiently high to discourage imitation by opportunistically oriented leaders.

The credibility of authenticity-related communication therefore depends not merely on differences in signaling costs but on their relationship with the strategic benefits associated with obtaining employee trust.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

4.3. Pooling Equilibria and the Loss of Signal Informativeness

A pooling equilibrium occurs when both leader types adopt identical communication strategies. Under such conditions, the observed signal does not enable employees to distinguish between relationally and opportunistically oriented leaders.

4.3.1. Pooling Equilibrium on the High-Authenticity Signal

Consider a strategic configuration in which both leader types select the high-authenticity signal:

$$\sigma_R(s_H) = \sigma_O(s_H) = 1 \tag{34}$$

Applying Bayes' rule yields:

$$\mu(s_H) = \frac{p}{p + (1 - p)} = p \tag{35}$$

The posterior belief remains identical to the prior belief. Consequently, the high-authenticity signal provides no additional information regarding the leader's underlying relational orientation.

Assume that the employee trusts after observing s_H and withholds trust following the off-equilibrium signal s_L. Such behavior is sequentially rational when:

$$p \geq \mu^*, \mu(s_L) \leq \mu^* \tag{36}$$

with appropriate trust decisions at equality.

Neither leader type has an incentive to deviate when:

$$c_R(\alpha) \leq W_R, c_O(\alpha) \leq W_O \tag{37}$$

Equations (36) and (37) establish sufficient conditions for a high-signal pooling Perfect Bayesian Equilibrium.

Under this configuration, authenticity-related communication remains observable and potentially persuasive but loses its capacity to distinguish between heterogeneous relational orientations.

4.3.2. Pooling Equilibrium on the Conventional Signal

An alternative pooling configuration occurs when both leader types select the conventional signal:

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

$$\sigma_R(s_L) = \sigma_O(s_L) = 1 \tag{38}$$

Bayesian consistency requires:

$$\mu(s_L) = p \tag{39}$$

Assuming non-negative incremental signaling costs and sufficiently pessimistic beliefs following the off-equilibrium high-authenticity signal, conventional-signal pooling can be supported as a Perfect Bayesian Equilibrium.

The employee's trust decision depends on the prior probability:

$$t^*(s_L) = \begin{cases} 1, & p > \mu^* \\ 0, & p < \mu^* \\ \text{either action or mixing}, & p = \mu^* \end{cases} \tag{40}$$

Unlike a separating equilibrium, pooling configurations produce trust decisions based on prior expectations rather than informative communication.

Proposition 3 (P3) – Authenticity signaling redundancy.

When relationally and opportunistically oriented leaders adopt identical communication strategies, authenticity-related signals lose their informational value. In such circumstances, interpersonal trust decisions depend primarily on employees' prior beliefs and anticipated relational consequences.

The existence of multiple pooling configurations also indicates that the disappearance of a separating equilibrium does not necessarily produce a unique alternative equilibrium.

4.4. Semi-Separating Equilibrium and Strategic Imitation

An intermediate equilibrium configuration arises when relationally oriented leaders consistently produce high-authenticity signals, while opportunistically oriented leaders probabilistically imitate them.

Let:

$$\sigma_R(s_H) = 1 \tag{41}$$

$$\sigma_O(s_H) = \lambda,\ 0 < \lambda < 1 \tag{42}$$

where λ denotes the probability of strategic imitation by opportunistically oriented leaders.

Following observation of the high-authenticity signal, the employee's posterior belief becomes:

$$\mu(s_H) = \frac{p}{p + (1 - p)\lambda} \tag{43}$$

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

The conventional signal identifies the opportunistic leader type:

$$\mu(s_L) = 0 \tag{44}$$

The effect of strategic imitation on posterior beliefs is obtained by differentiating Equation (43):

$$\frac{\partial \mu(s_H)}{\partial \lambda} = -\frac{p(1-p)}{[p + (1-p)\lambda]^2} < 0 \tag{45}$$

Equation (45) demonstrates that increasing imitation reduces the informational credibility of high-authenticity signals.

For an interior semi-separating equilibrium involving employee randomization, the employee must remain indifferent between extending and withholding trust after observing s_H:

$$\mu(s_H) = \mu^* \tag{46}$$

Combining Equations (43) and (46) yields:

$$\frac{p}{p + (1-p)\lambda^*} = \mu^*$$

Solving for the equilibrium imitation probability:

$$\boxed{\lambda^* = \frac{p(1-\mu^*)}{(1-p)\mu^*}} \tag{47}$$

Substituting the trust threshold from Equation (23):

$$\boxed{\lambda^* = \frac{pB}{(1-p)K}} \tag{48}$$

An interior imitation probability exists when:

$$0 < p < \mu^* \tag{49}$$

Let τ denote the probability that the employee extends trust after observing the high-authenticity signal. The opportunistic leader must be indifferent between selecting the conventional signal and strategically imitating the relational leader:

$$\tau W_O - c_O(\alpha) = 0 \tag{50}$$

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Therefore:

$$\tau^*(\alpha) = \frac{c_O(\alpha)}{W_O} \tag{51}$$

For employee randomization to remain interior:

$$0 < c_O(\alpha) < W_O \tag{52}$$

The relational leader's incentive compatibility condition becomes:

$$\tau^*(\alpha) W_R - c_R(\alpha) \geq 0 \tag{53}$$

Substituting Equation (51):

$$\frac{c_R(\alpha)}{W_R} \leq \frac{c_O(\alpha)}{W_O} \tag{54}$$

Equations (49), (52), and (54), together with non-negative signaling costs, establish the conditions supporting the proposed semi-separating equilibrium.

Proposition 4 (P4) – Strategic imitation and probabilistic trust.

Under an interior semi-separating equilibrium, opportunistically oriented leaders probabilistically imitate authenticity-related signals, while employees probabilistically extend trust. Holding continuation payoffs constant, a reduction in opportunistic signaling costs decreases the equilibrium probability of employee trust.

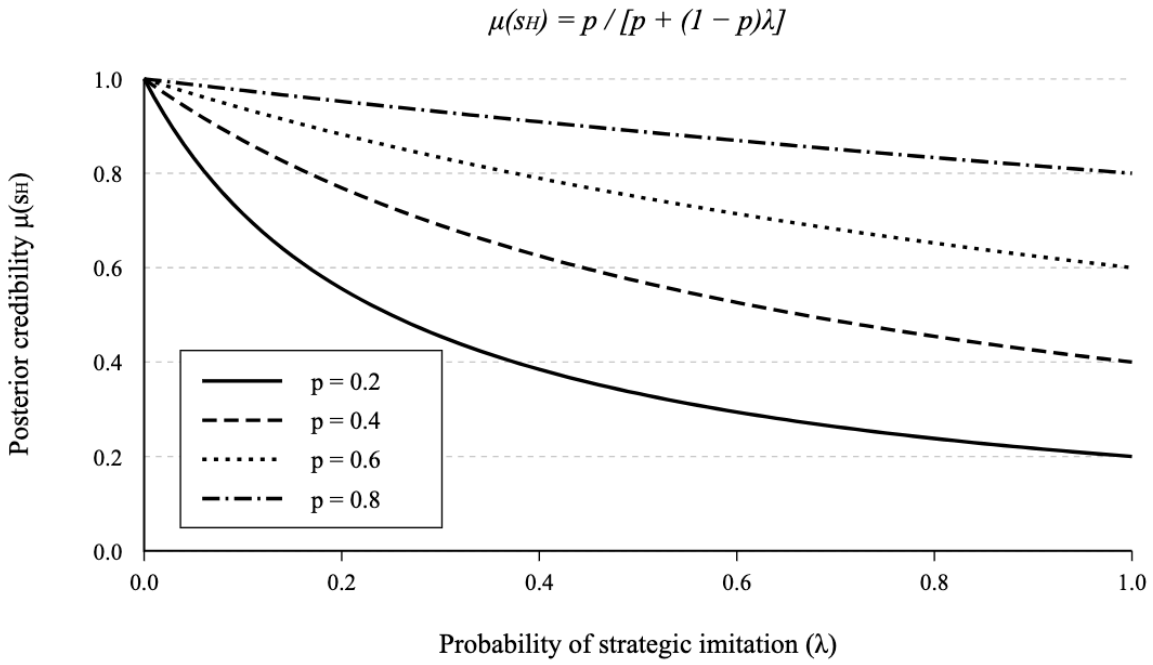
Specifically:

$$\frac{d\tau^*(\alpha)}{d\alpha} = \frac{c'_O(\alpha)}{W_O} \leq 0 \tag{55}$$

Notably, within this equilibrium, the imitation probability λ* depends on prior beliefs and employee payoffs rather than directly on AI mediation. Changes in signaling costs are accommodated through employee trust probabilities, provided the equilibrium conditions remain satisfied.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Figure 3. Effect of Strategic Imitation on the Posterior Credibility of Authenticity Signals



Source: Authors' analytical representation based on Equation (43).
The curves represent hypothetical prior probabilities rather than empirical observations.

4.5. Formalization of the AI-Induced Authenticity Signal Dilution Effect (AISDE)

To investigate the technological mechanism underlying AISDE, the incremental signaling cost functions are expressed through a linear specification:

$$c_R(\alpha) = a_R - b_R\alpha \tag{56}$$

$$c_O(\alpha) = a_O - b_O\alpha \tag{57}$$

subject to:

$$a_O > a_R > 0$$

$$0 \leq b_R < b_O,\ b_i \leq a_i \tag{58}$$

where a_i represents the incremental signaling cost without AI assistance and b_i represents the sensitivity of signaling costs to technological mediation.

These restrictions ensure non-negative signaling costs throughout the technological mediation interval.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

The cost differential becomes:

$$\Delta c(\alpha) = (a_O - a_R) - (b_O - b_R)\alpha \tag{59}$$

Differentiating Equation (59):

$$\boxed{\frac{d\Delta c(\alpha)}{d\alpha} = -(b_O - b_R) < 0} \tag{60}$$

Equation (60) formalizes the cost-compression mechanism assumed in A5.

From Equation (33), a credible separating equilibrium requires:

$$a_O - b_O\alpha \geq W_O \tag{61}$$

Solving for the degree of technological mediation yields the critical AI mediation threshold:

$$\boxed{\alpha^* = \frac{a_O - W_O}{b_O}} \tag{62}$$

An interior threshold exists when:

$$a_O - b_O < W_O < a_O \tag{63}$$

Provided the relational leader's incentive compatibility condition remains satisfied, the separating equilibrium is feasible for:

$$\alpha \leq \alpha^*$$

and becomes infeasible when:

$$\boxed{\alpha > \alpha^*} \tag{64}$$

Beyond this threshold, the opportunistic leader obtains a strictly positive payoff from imitating the high-authenticity signal under the separating employee strategy.

Proposition 5 (P5) – AI-Induced Authenticity Signal Dilution Effect.

If AI-mediated communication disproportionately reduces the incremental signaling costs of opportunistically oriented leaders, and such costs fall below the strategic benefits associated with obtaining employee trust, the separating equilibrium loses incentive compatibility.

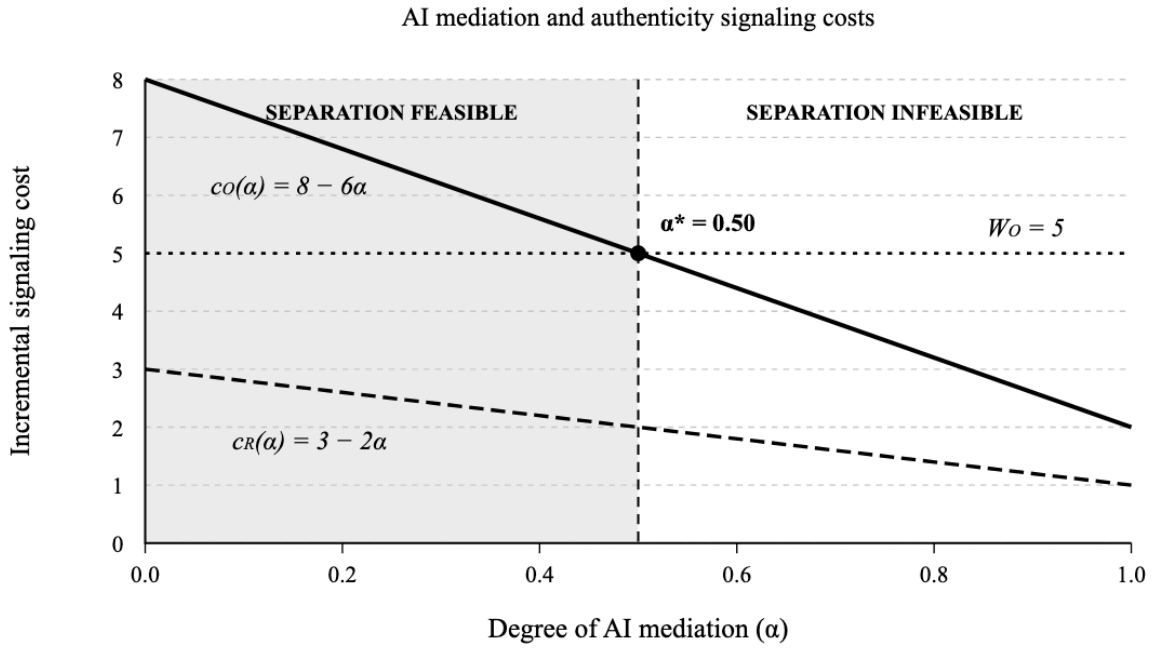
AISDE is consequently conceptualized through two analytically distinct mechanisms:

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

- **Technological cost compression:** AI reduces the differential costs of producing authenticity-related signals.
- **Strategic signal dilution:** Cost reduction crosses the incentive compatibility threshold, allowing opportunistic leaders to imitate previously distinguishing signals.

The first mechanism is a modeling assumption expressed through Equation (60). The second is a deductive result established through Equations (61)–(64).

Figure 4. The AI-Induced Authenticity Signal Dilution Effect (AISDE)



Illustrative parameters: $a_O = 8$, $b_O = 6$, $a_R = 3$, $b_R = 2$, $W_O = 5$, $W_R = 4$

Source: Authors' theoretical development based on Equations (56)–(64).

The critical technological mediation threshold is not a universal characteristic of AI adoption. Rather, it represents a context-dependent function of signaling costs and strategic incentives.

4.6. Numerical Illustration and Comparative Equilibrium Analysis

To illustrate the analytical implications of the Leadership Authenticity Signaling Game (LASG), a hypothetical parameter configuration is introduced. The selected values serve exclusively as a numerical demonstration of the theoretical model and do not represent empirical estimates or observed organizational behavior.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Table 3. Illustrative Parameter Configuration of the LASG Model

Parameter	Description	Value
p	Prior probability of a relational leader	0.40
B	Employee benefit from reciprocity	4
K	Employee loss from opportunistic behavior	6
W_r	Relational leader's optimal continuation payoff	4
W_o	Opportunistic leader's optimal continuation payoff	5
a_r	Baseline incremental signaling cost: relational leader	3
a_o	Baseline incremental signaling cost: opportunistic leader	8
b_r	AI-induced cost sensitivity: relational leader	2
b_o	AI-induced cost sensitivity: opportunistic leader	6

Source: Authors' hypothetical parameter specification.

Substituting the selected parameters into Equation (23), the employee trust threshold becomes:

$$\mu^* = \frac{K}{B+K} = \frac{6}{4+6} = 0.60 \tag{65}$$

The employee requires a posterior probability of at least 60% to consider extending trust strategically rational under this parameter configuration.

The incremental signaling cost functions are consequently specified as:

$$c_R(\alpha) = 3 - 2\alpha$$

$$c_O(\alpha) = 8 - 6\alpha$$

$$\Delta c(\alpha) = 5 - 4\alpha$$

Applying Equation (62), the critical AI mediation threshold becomes:

$$\alpha^* = \frac{a_O - W_O}{b_O} = \frac{8-5}{6} = 0.50 \tag{66}$$

For technological mediation levels exceeding 0.50, the opportunistic leader's incremental signaling cost falls below the strategic benefit of obtaining trust. The separating equilibrium therefore loses incentive compatibility.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Since the prior probability satisfies $p = 0.40 < \mu^* = 0.60$, an interior semi-separating equilibrium may emerge when the remaining incentive compatibility conditions are satisfied.

Applying Equation (48), the equilibrium imitation probability becomes:

$$\lambda^* = \frac{pB}{(1-p)K} = \frac{0.40 \times 4}{0.60 \times 6} \approx 0.4444 \quad (67)$$

Within the specified interior semi-separating equilibrium, opportunistically oriented leaders imitate high-authenticity signals with a probability of approximately 44.44%.

The corresponding equilibrium probability of employee trust conditional on observing the high-authenticity signal is obtained from Equation (51):

$$\tau^*(\alpha) = \frac{c_O(\alpha)}{W_O} = \frac{8 - 6\alpha}{5} \quad (68)$$

Table 4 presents the comparative equilibrium implications associated with alternative degrees of AI-mediated leadership communication.

Table 4. Comparative Equilibrium Analysis across Different Degrees of AI Mediation

AI (α)	Cost $c_r(\alpha)$	Cost $c_o(\alpha)$	Cost gap $\Delta c(\alpha)$	Separating equilibrium feasible	Interior semi-separating feasible	Conditional employee trust
0.00	3.00	8.00	5.00	Yes	No	100%
0.20	2.60	6.80	4.20	Yes	No	100%
0.40	2.20	5.60	3.40	Yes	No	100%
0.50	2.00	5.00	3.00	Yes*	No	100%
0.60	1.80	4.40	2.60	No	Yes	88%
0.80	1.40	3.20	1.80	No	Yes	64%
1.00	1.00	2.00	1.00	No	Yes	40%

** Boundary condition: the opportunistic leader is indifferent between signaling strategies.*

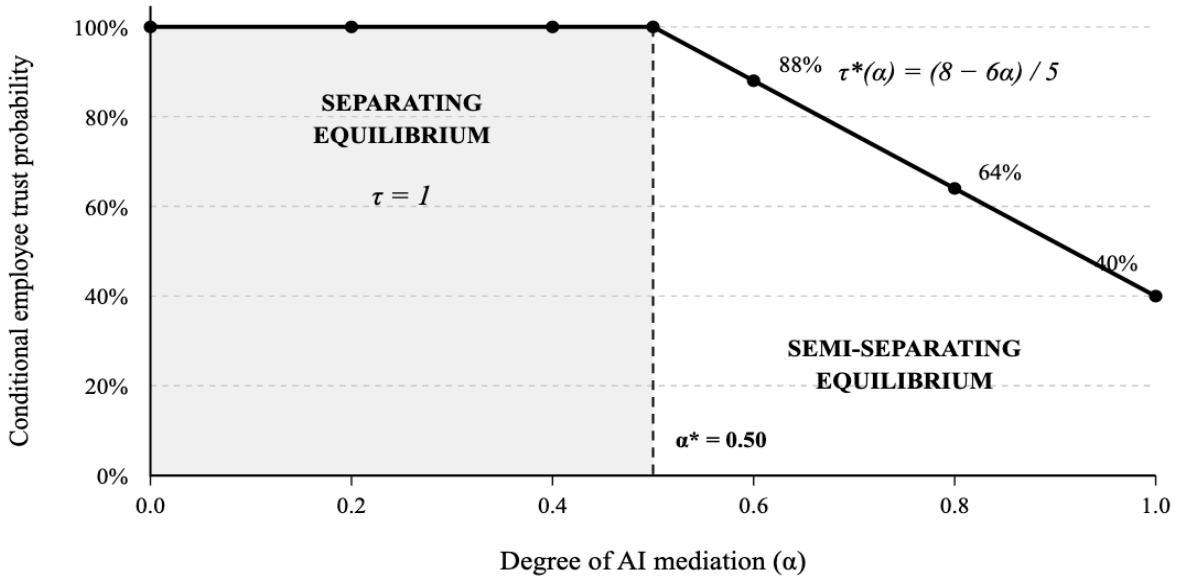
Source: Authors' analytical calculations based on the parameters presented in Table 3.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

The numerical results illustrate three distinct strategic configurations:

1. For $\alpha < 0.50$, separating strategies remain incentive-compatible.
2. At $\alpha = 0.50$, the opportunistic leader reaches the indifference boundary.
3. For $\alpha > 0.50$, the selected semi-separating equilibrium becomes feasible, and conditional employee trust decreases as AI mediation increases.

Figure 5. AI Mediation and Equilibrium Conditional Employee Trust



Source: Authors' analytical representation based on Equation (68) and Table 4.
The selected equilibrium configurations and parameter values are illustrative.

The numerical illustration demonstrates that technological reductions in signaling costs may undermine the credibility of authenticity-related communication and reduce conditional employee trust within the selected equilibrium configurations.

Importantly, the equilibrium imitation probability remains constant within the specified semi-separating regime, while the conditional probability of employee trust adjusts to changes in the opportunistic leader's signaling costs.

These findings provide a numerical illustration of the AISDE mechanism rather than empirical evidence of its existence or magnitude in actual organizations.

4.7. Accountability-Based Restoration of Authenticity Signal Credibility

Although AI-mediated communication may reduce the informational credibility of authenticity-related signals, organizational mechanisms that introduce verifiable commitments and behavioral accountability may restore the strategic conditions necessary for credible differentiation between leader types.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

To incorporate this possibility into LASG, an additional parameter is introduced:

$$z \in [0, 1]$$

where z represents the intensity of verifiable relational commitment associated with the high-authenticity signal.

Such commitments may include documented managerial responsibilities, observable follow-up actions, measurable behavioral obligations, and mechanisms through which employees can verify consistency between communication and conduct.

The incremental signaling cost function is extended to incorporate the additional cost associated with verifiable commitment:

$$\tilde{c}_i(\alpha, z) = c_i(\alpha) + k_i z \tag{69}$$

where $k_i \geq 0$ represents the type-specific incremental cost of satisfying the verifiable commitment requirements.

Under this extended specification, the incentive compatibility condition for the relationally oriented leader becomes:

$$c_R(\alpha) + k_R z \leq W_R \tag{70}$$

The corresponding incentive compatibility condition for the opportunistically oriented leader becomes:

$$c_O(\alpha) + k_O z \geq W_O \tag{71}$$

Equations (70) and (71) indicate that accountability mechanisms may discourage strategic imitation by increasing its associated costs, provided that the additional commitment requirements do not simultaneously discourage relationally oriented leaders from producing the high-authenticity signal.

Assuming strictly positive commitment cost coefficients, the minimum intensity of verifiable commitment required to discourage opportunistic imitation is:

$$z_{min} = max \left(0, \frac{W_O - c_O(\alpha)}{k_O} \right) \tag{72}$$

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Similarly, the maximum commitment intensity compatible with the relational leader's willingness to produce the high-authenticity signal is:

$$z_{max} = min\left(1, \frac{W_R - c_R(\alpha)}{k_R}\right) \tag{73}$$

Consequently, the restoration of a separating equilibrium through verifiable commitment is feasible if and only if:

$$z_{min} \leq z_{max} \tag{74}$$

This condition establishes the existence of a feasible accountability interval within which opportunistic imitation becomes unattractive without eliminating the incentives of relationally oriented leaders to communicate authentically.

Under the linear signaling cost specification introduced in Equations (56) and (57), the opportunistic leader's incentive compatibility condition becomes:

$$a_O - b_O\alpha + k_O z \geq W_O$$

Solving for the critical degree of AI mediation yields an accountability-adjusted technological mediation threshold:

$$\alpha^*(z) = \frac{a_O + k_O z - W_O}{b_O} \tag{75}$$

Equation (75) demonstrates how verifiable commitment modifies the technological boundary beyond which opportunistic imitation becomes strategically profitable.

Differentiating Equation (75) with respect to commitment intensity yields:

$$\frac{d\alpha^*(z)}{dz} = \frac{k_O}{b_O} > 0 \tag{76}$$

For strictly positive opportunistic commitment costs, Equation (76) establishes that increasing verifiable commitment raises the critical AI mediation threshold, provided that the relational leader's incentive compatibility condition continues to hold.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Proposition 6 (P6) – Accountability-based restoration of signal credibility.

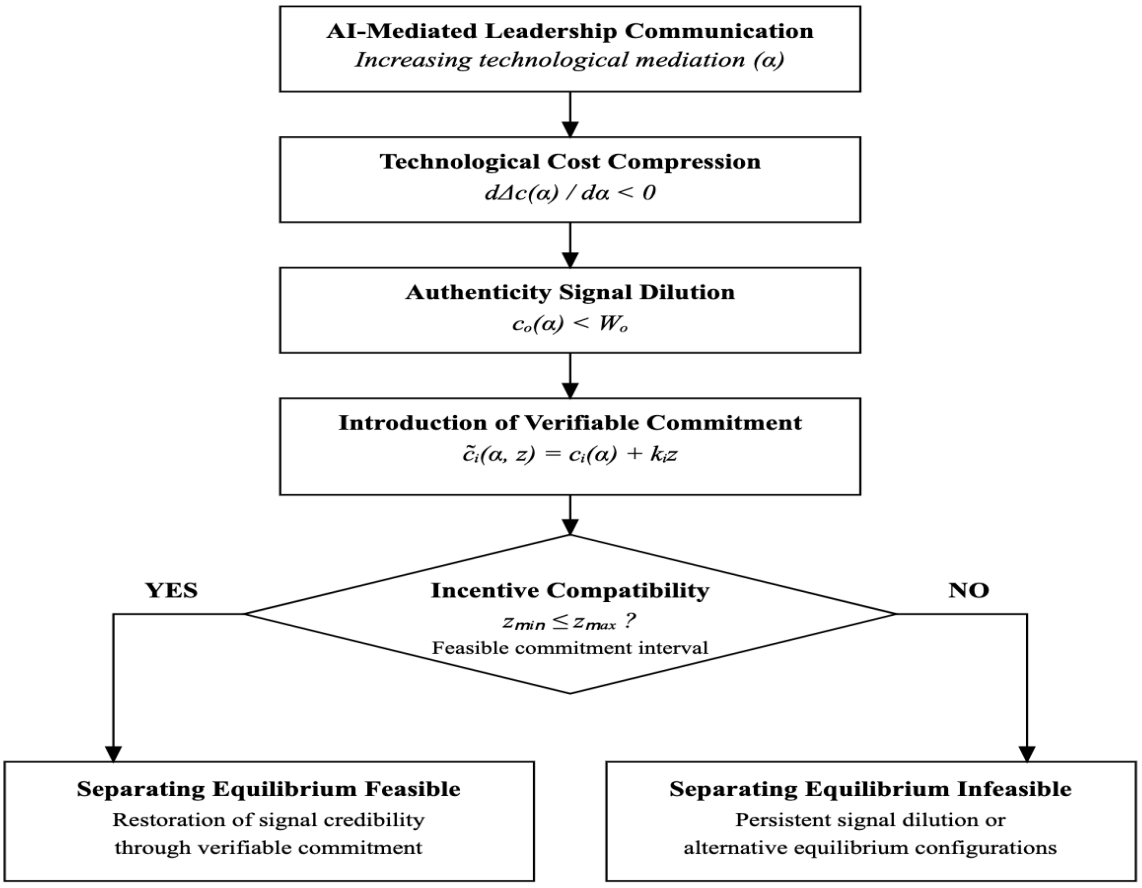
Verifiable relational commitments and accountability mechanisms may restore the informational credibility of authenticity-related signals when their associated incremental costs discourage opportunistic imitation while preserving the strategic incentives of relationally oriented leaders.

This result suggests that organizational responses to AISDE need not necessarily involve restricting the use of artificial intelligence. Instead, organizations may strengthen the connection between leadership communication and observable behavioral commitments.

Transparency regarding AI involvement may provide complementary information. However, within the baseline LASG specification, where technological mediation is common knowledge, disclosure alone does not necessarily distinguish between relationally and opportunistically oriented leaders.

The restoration mechanism therefore depends on the strategic consequences of verifiable commitments rather than exclusively on the perceived authenticity of AI-mediated leadership communication.

Figure 6. Accountability-Based Restoration of Authenticity Signal Credibility



Source: Authors' theoretical development based on Equations (69)–(76).

The proposed mechanism illustrates how accountability may counteract the strategic consequences of AI-induced authenticity signal dilution.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

4.8. Synthesis of Theoretical Results

The formal development of the Leadership Authenticity Signaling Game (LASG) provides an integrated analytical framework for examining the strategic relationship between AI-mediated leadership communication, authenticity-related signaling, reciprocity, and interpersonal trust.

The equilibrium analysis identifies six theoretical propositions that collectively explain the conditions under which authenticity-related communication retains, loses, or potentially recovers its informational credibility.

These propositions establish the analytical foundations of the proposed AI-Induced Authenticity Signal Dilution Effect (AISDE), while identifying the strategic conditions under which organizational accountability mechanisms may counteract the consequences of technological cost compression.

Table 5 synthesizes the principal theoretical propositions, their mathematical foundations, and their implications for authenticity-related communication and interpersonal trust in AI-mediated organizational environments.

Table 5. Synthesis of the Theoretical Propositions Derived from the LASG Model

No.	Theoretical proposition	Mathematical foundation	Theoretical implication
P1	**Vulnerability-dependent trust threshold**	$\mu^* = \frac{K}{B+K}$ $\partial\mu^*/\partial K > 0;\ \partial\mu^*/\partial B < 0$	Greater vulnerability increases the minimum posterior belief required for employees to rationally extend interpersonal trust.
P2	**Incentive-compatible authenticity signaling**	$c_r(\alpha) \le W_r$ $c_o(\alpha) \ge W_o$	Authenticity signals distinguish leader types when signaling costs discourage opportunistic imitation while preserving relational leaders' incentives.
P3	**Authenticity signaling redundancy**	$\sigma_r(s) = \sigma_o(s) = 1$ $\mu(s) = p$	Identical signaling strategies eliminate the informational contribution of the observed signal, leaving trust decisions dependent on prior beliefs.
P4	**Strategic imitation and probabilistic trust**	$\lambda^* = \frac{pB}{(1-p)K}$ $\tau^* = \frac{c_o(\alpha)}{W_o}$	Within the selected mixed equilibrium, lower imitation costs reduce conditional employee trust, holding continuation payoffs constant.
P5	**AI-Induced Authenticity Signal Dilution Effect (AISDE)**	$\alpha^* = \frac{a_o - W_o}{b_o}$ $\alpha > \alpha^* \Rightarrow$ *separation infeasible*	AI-induced cost reductions may eliminate the incentive compatibility of separating strategies beyond a context-dependent technological threshold.
P6	**Accountability-based restoration of signal credibility**	$z_{min} \le z_{max}$ $d\alpha^*(z)/dz = k_o/b_o > 0$	Verifiable commitments may restore separation when accountability costs discourage imitation without deterring relationally oriented leaders.

Source: Authors' theoretical development based on the formal analysis of LASG.
Note: All propositions are conditional on the corresponding modeling and equilibrium assumptions.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Collectively, the theoretical results indicate that the informational credibility of authenticity-related leadership communication depends on the interaction between employee vulnerability, heterogeneous leader preferences, signaling costs, and the strategic benefits associated with obtaining interpersonal trust.

The principal contribution of LASG consists in demonstrating that technological mediation may alter the strategic conditions under which authenticity-related communication functions as an informative signal of a leader's underlying relational orientation.

Importantly, the proposed AISDE mechanism does not imply that AI-mediated communication necessarily produces inauthentic leadership. Rather, it identifies the conditions under which technological assistance may reduce the capacity of communication to differentiate between heterogeneous leader types.

The formal analysis further distinguishes the informational credibility of a signal from the probability of extending trust and the probability of actual reciprocity. These outcomes are strategically interconnected but should not be interpreted as conceptually or mathematically interchangeable.

The accountability extension demonstrates that the strategic consequences of technological cost compression may be mitigated through verifiable commitments that reconnect authenticity-related communication with observable managerial behavior and relational responsibility.

Nevertheless, the existence of multiple equilibrium configurations prevents the identification of a universally determined transition between separating, pooling, and semi-separating strategies.

The numerical illustrations should therefore be interpreted as demonstrations of the internal analytical implications of the proposed model rather than as empirical predictions of actual organizational behavior.

These findings establish the theoretical foundation for discussing the implications of AI-mediated communication for authentic leadership theory, interpersonal trust, managerial accountability, leadership development, and future empirical investigations of human–AI collaboration in organizational environments.

5. DISCUSSION

5.1. Theoretical Contributions to Authentic Leadership and Interpersonal Trust

The Leadership Authenticity Signaling Game (LASG) extends the conceptual understanding of authentic leadership by introducing a strategic perspective on the relationship between communication, relational orientation, reciprocity, and interpersonal trust. While existing authentic leadership research emphasizes self-awareness, relational transparency, and behavioral consistency [3–6,16–20], the proposed model examines the conditions under which employees can infer underlying leadership orientations from observable communication.

Propositions P1 and P2 suggest that the credibility of authenticity-related signals depends on employee vulnerability and the incentive structures governing leadership behavior. This interpretation

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

complements organizational trust theory [7], social exchange theory [21–23], and signaling theory [12,13] by explicitly incorporating information asymmetry and strategic imitation.

The theoretical contribution of LASG lies in distinguishing three interconnected dimensions: the perceived authenticity of leadership communication, its informational credibility, and the actual reciprocity of managerial behavior. Authenticity-related communication may remain emotionally persuasive while becoming less informative about the leader's underlying relational orientation.

This distinction is particularly relevant in AI-mediated organizational environments, where linguistic sophistication may increasingly reflect technological capabilities rather than individual relational investment.

5.2. The AI-Induced Authenticity Signal Dilution Effect

The proposed AI-Induced Authenticity Signal Dilution Effect (AISDE) provides a formal explanation of how technological assistance may alter the informational properties of leadership communication. This mechanism complements emerging research concerning AI-mediated communication, perceived empathy, message authorship, and interpersonal trust [1,8–11,15,28,29].

Previous research suggests that AI-generated communication can produce positive interpersonal experiences while knowledge of artificial authorship may diminish their perceived relational value [9,29].

Semantic Scholar+1
AISDE introduces a complementary explanation based on strategic incentives rather than exclusively on psychological reactions to AI involvement.

Propositions P3–P5 demonstrate that technological reductions in signaling costs may create conditions under which opportunistically oriented leaders can imitate previously distinguishing authenticity signals. Consequently, employees may become less capable of differentiating between relational commitment and strategically produced emotional expression.

Importantly, the model does not establish that increasing AI mediation necessarily reduces the aggregate probability of interpersonal trust. Conditional trust decreases within the specified interior semi-separating equilibrium, whereas unconditional trust additionally depends on the distribution of observable signals. Similarly, equilibrium multiplicity prevents the identification of a universally determined transition between strategic configurations.

The leadership authenticity dilemma therefore emerges from the potential tension between technological communication efficiency and the preservation of informative relational signals.

5.3. Managerial and Organizational Implications

The proposed framework contributes to the automation–augmentation debate in management [2] by suggesting that the organizational value of AI-mediated leadership communication should be evaluated beyond conventional measures of efficiency and productivity.

From a managerial perspective, AI assistance may remain beneficial when it improves communication without weakening behavioral responsibility. However, excessive reliance on technologically generated emotional expression may create strategic conditions that undermine the informational value of authenticity-related communication.

Proposition P6 suggests that organizations may counteract this mechanism through appropriately designed accountability structures. These may include documented managerial commitments, observable follow-up actions, employee participation in evaluating leadership conduct, and explicit responsibility for AI-assisted communication.

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

Such mechanisms should preserve the advantages of technological augmentation while strengthening the relationship between managerial communication and subsequent behavior.

Transparency concerning AI involvement may complement these practices, particularly given emerging research on AI disclosure and trust [11,15]. Nevertheless, transparency should not be considered a sufficient substitute for behavioral consistency. Within LASG, the restoration of signal credibility depends on whether accountability requirements effectively alter the incentives associated with strategic imitation.

5.4. Implications for Leadership Development and Management Education

The LASG framework also offers potential educational applications in leadership development, managerial ethics, organizational behavior, and responsible AI adoption.

Rather than approaching authentic leadership exclusively through normative principles or individual competencies, management education may incorporate strategic simulations illustrating how communication choices, heterogeneous preferences, technological capabilities, and employee vulnerability influence interpersonal outcomes.

The integration of game-based learning approaches into educational environments provides a complementary pedagogical perspective [30]. LASG could therefore serve as the theoretical foundation for interactive managerial simulations in which participants assume different organizational roles, evaluate signaling strategies, and experience the consequences of alternative trust and reciprocity decisions.

Such educational applications may facilitate the development of strategic reasoning, ethical awareness, and a more comprehensive understanding of the relational responsibilities associated with AI-assisted leadership.

5.5. Research Limitations and Future Research Directions

Several limitations should be acknowledged when interpreting the theoretical results.

First, LASG represents leadership heterogeneity through two stylized behavioral types. Actual managerial orientations are likely to be multidimensional, context-dependent, and subject to change over time.

Second, the baseline model assumes rational decision-making, Bayesian belief updating, and common knowledge of technological mediation. Behavioral biases, emotional responses, organizational cultures, and incomplete information regarding AI involvement may substantially influence actual employee decisions.

Third, the differential technological impact on signaling costs constitutes a modeling assumption requiring empirical investigation. AI may also reduce communication barriers for genuinely relational leaders, potentially strengthening their ability to express authentic intentions.

Fourth, the accountability extension treats verifiable commitment intensity as an observable parameter rather than an endogenous strategic choice. Alternative accountability mechanisms may also affect continuation payoffs and actual reciprocal behavior.

Fifth, the proposed equilibrium configurations are not necessarily unique or exhaustive. Future theoretical extensions could employ equilibrium refinement techniques, repeated games, evolutionary game theory, and dynamic reputation models to investigate the stability of authenticity-related communication strategies.

Future empirical research could employ experimental trust games adapted to leadership communication [14,24,26], manipulating AI mediation, employee vulnerability, disclosure, and verifiable managerial commitment. Longitudinal organizational studies could subsequently examine

DOI: https://doi.org/10.65222/VIRAL.2026.10.56.76

whether repeated interactions and observable behavioral consistency moderate the proposed AISDE mechanism.

Ultimately, combining formal modeling with behavioral experiments and organizational field research would enable the evaluation, refinement, and potential empirical validation of LASG and AISDE.

6. CONCLUSION

The increasing integration of artificial intelligence into managerial communication raises important theoretical and practical questions concerning the relationship between leadership authenticity, reciprocity, and interpersonal trust. This study addressed these questions by developing the **Leadership Authenticity Signaling Game (LASG)**, a formal framework integrating authentic leadership theory, signaling theory, trust games, and social exchange theory.

The theoretical analysis demonstrated that the informational credibility of authenticity-related communication depends on the interaction between heterogeneous leader preferences, signaling costs, employee vulnerability, and strategic incentives. Through the examination of separating, pooling, and semi-separating equilibria, the study established six theoretical propositions explaining the conditions under which authenticity-related signals may retain, lose, or recover their informational value.

A central theoretical contribution is the proposed **AI-Induced Authenticity Signal Dilution Effect (AISDE)**, which explains how technological reductions in signaling costs may undermine the incentive compatibility of previously credible authenticity signals. Nevertheless, the model does not suggest that AI-mediated communication is inherently inauthentic or necessarily detrimental to interpersonal trust. Instead, its relational consequences depend on the strategic conditions governing communication and subsequent managerial behavior.

The accountability extension further demonstrated that verifiable relational commitments may restore the feasibility of credible authenticity signaling without necessarily restricting technological adoption. This finding emphasizes the importance of connecting AI-assisted communication with observable managerial conduct, behavioral consistency, and relational responsibility.

Although the proposed framework remains theoretical and requires empirical validation, it establishes an analytical foundation for future research on authentic leadership in AI-mediated organizational environments. Its potential applications extend to managerial decision-making, organizational governance, leadership development, and management education.

Ultimately, the **leadership authenticity dilemma** concerns not merely whether artificial intelligence participates in managerial communication, but whether such communication remains a credible indicator of human intentions and reciprocal behavior. As artificial intelligence increasingly enhances the capacity to communicate authenticity, the strategic and organizational importance of demonstrating authenticity through accountable human action may become increasingly significant.

ACKNOWLEDGMENTS

The authors acknowledge the contributions of the scientific community whose research in authentic leadership, interpersonal trust, game theory, and artificial intelligence provided the theoretical foundations for this study. The authors also acknowledge the use of chatGPT as a supportive tool during the conceptual development, mathematical formulation, manuscript preparation, and linguistic refinement of this research. The authors retain full responsibility for the theoretical framework, mathematical reasoning, interpretation of results, and scientific integrity of the manuscript.

7. REFERENCES

1. Hancock JT, Naaman M, Levy K. AI-mediated communication: Definition, research agenda, and ethical considerations. J Comput Mediat Commun. 2020;25(1):89–100. https://doi.org/10.1093/jcmc/zmz022

2. Raisch S, Krakowski S. Artificial intelligence and management: The automation–augmentation paradox. Acad Manage Rev. 2021;46(1):192–210. https://doi.org/10.5465/amr.2018.0072

3. Gardner WL, Cogliser CC, Davis KM, Dickens MP. Authentic leadership: A review of the literature and research agenda. Leadersh Q. 2011;22(6):1120–1145. https://doi.org/10.1016/j.leaqua.2011.09.007

4. Alvesson M, Einola K. Warning for excessive positivity: Authentic leadership and other traps in leadership studies. Leadersh Q. 2019;30(4):383–395.

5. Algera PM, Lips-Wiersma M. Radical authentic leadership: Co-creating the conditions under which all members of the organization can be authentic. Leadersh Q. 2012;23(1):118–131.

6. Ibarra H. The authenticity paradox. Harv Bus Rev. 2015;93(1–2):53–59.

7. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Acad Manage Rev. 1995;20(3):709–734. https://doi.org/10.5465/amr.1995.9508080335

8. Hohenstein J, Kizilcec RF, DiFranzo D, et al. Artificial intelligence in communication impacts language and social relationships. Sci Rep. 2023;13:5487. https://doi.org/10.1038/s41598-023-30938-9

9. Yin Y, Jia N, Wakslak CJ. AI can help people feel heard, but an AI label diminishes this impact. Proc Natl Acad Sci USA. 2024;121(14):e2319112121. https://doi.org/10.1073/pnas.2319112121

10. Jakesch M, Hancock JT, Naaman M. Human heuristics for AI-generated language are flawed. Proc Natl Acad Sci USA. 2023;120(11):e2208839120. https://doi.org/10.1073/pnas.2208839120

11. Schilke O, Reimann M. The transparency dilemma: How AI disclosure erodes trust. Organ Behav Hum Decis Process. 2025;188:104405. https://doi.org/10.1016/j.obhdp.2025.104405

12. Spence M. Job market signaling. Q J Econ. 1973;87(3):355–374. https://doi.org/10.2307/1882010

13. Connelly BL, Certo ST, Ireland RD, Reutzel CR. Signaling theory: A review and assessment. J Manage. 2011;37(1):39–67. https://doi.org/10.1177/0149206310388419

14. Berg J, Dickhaut J, McCabe K. Trust, reciprocity, and social history. Games Econ Behav. 1995;10(1):122–142. https://doi.org/10.1006/game.1995.1027

15. Baum HM, Gerpott FH, Kerschreiter R. (A)I Did That: Toward a model of leaders' AI use and disclosure effects on employees. J Leadersh Organ Stud. 2026. https://doi.org/10.1177/15480518261469227

16. Gardner WL, Karam EP, Alvesson M, Einola K. Authentic leadership theory: The case for and against. Leadersh Q. 2021;32(6):101495. https://doi.org/10.1016/j.leaqua.2021.101495

17. Banks GC, McCauley KD, Gardner WL, Guler CE. A meta-analytic review of authentic and transformational leadership: A test for redundancy. Leadersh Q. 2016;27(4):634–652. https://doi.org/10.1016/j.leaqua.2016.02.006

18. Shamir B, Eilam-Shamir G. "What's your story?" A life-stories approach to authentic leadership development. In: Leadership Now: Reflections on the Legacy of Boas Shamir. Emerald Publishing; 2018.

19. Liu H, Cutcher L, Grant D. Doing authenticity: The gendered construction of authentic leadership. Gend Work Organ. 2015;22(3):237–255.

20. Ayaz O, Ozbilgin MF, Samdanis M, Torunoğlu Tinay D. Authenticity and atypicality in leadership: Can an atypical leader afford to be authentic? Int J Manag Rev. 2024;26(2):312–328.

21. Cropanzano R, Anthony EL, Daniels SR, Hall AV. Social exchange theory: A critical review with theoretical remedies. Acad Manage Ann. 2017;11(1):479–516. https://doi.org/10.5465/annals.2015.0099

22. Colquitt JA, LePine JA, Piccolo RF, Zapata CP, Rich BL. Explaining the justice–performance relationship: Trust as exchange deepener or trust as uncertainty reducer? J Appl Psychol. 2012;97(1):1–15. https://doi.org/10.1037/a0025208

23. Chen JK, Sriphon T. Authentic leadership, trust, and social exchange relationships under the influence of leader behavior. Sustainability. 2022;14(10):5883. https://doi.org/10.3390/su14105883

24. Smith A. Reciprocity effects in the trust game. Games. 2013;4(3):367–374. https://doi.org/10.3390/g4030367

25. Watson J. Strategy: An Introduction to Game Theory. 3rd ed. New York: W.W. Norton & Company; 2013.

26. Clark R. Trust and risk in games of partial information. Balt Int Yearb Cogn Log Commun. 2013;8:1–20. https://doi.org/10.4148/biyclc.v8i0.1073

27. Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. Acad Manage Ann. 2020;14(2):627–660. https://doi.org/10.5465/annals.2018.0057

28. Kirk CP, Givi J. The AI-authorship effect: Understanding authenticity, moral disgust, and consumer responses to AI-generated marketing communications. J Bus Res. 2025;186:114984. https://doi.org/10.1016/j.jbusres.2024.114984

29. Dorigoni G, Giardino PL. The illusion of empathy: Evaluating AI-generated outputs in moments that matter. Front Psychol. 2025. https://doi.org/10.3389/fpsyg.2025.1568911

30. Nguyen TD, Le TQ, Luong GT, Nguyen DT, Nguyen TT. A review of Applying Game-Based Learning and Learning Motivation: Game-Based Learning and Learning Motivation. EDUVIRAL [Internet]. 2025 Dec. 19 [cited 2026 Nov. 27];3(03):67-74. Available from: https://journalviral.org/home/article/view/30